Landscape Summary: Bias in Algorithmic Decision-Making

Landscape Summary:

Bias in Algorithmic Decision-Making

What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?

Authors:
Dr Michael Rovatsos (Edinburgh University)
Dr Brent Mittelstadt (Oxford Internet Institute)
Dr Ansgar Koene (University of Nottingham)

Centre for
Data Ethics
and Innovation

Executive Summary

As our societies become increasingly dependent on algorithms, so we are seeing our age-old prejudices, biases and implicit assumptions reflected back at us in digital form. But the algorithmic systems we use also have the potential to amplify, accentuate and systemise our biases on an unprecedented scale, all while presenting the appearance of objective, neutral arbiters.

This Landscape Summary draws together the literature and debates around algorithmic bias, the methods and strategies which may help to mitigate its impact, and explores four sectors in which this phenomenon is already starting to have real world consequences—financial services, local government, crime and justice, and recruitment. We identify a case study for each sector which may have significant consequences for individuals and groups in the UK: algorithmic loan redlining, child welfare, offender risk assessments, and CV sifting.

Through these case studies we have identified a number of key findings which are relevant for policymakers, regulators and other officials as they try to understand the socio-economic effects of algorithmic decision-making systems. These include:

- **There is no one generic approach to fairness, only alternative interpretations, which have implications for mitigating bias.** We cannot expect machines to reconcile these differences when society has not, and there will be trade-offs in any chosen approach. Attempts to mitigate algorithmic bias must therefore carefully consider what a fair outcome in any given context should look like, and develop strategies accordingly.

- **Legislation in the UK covers some, but not all, manifestations of algorithmic bias.** The Equality Act 2010 prohibits discrimination against people on the basis of certain protected characteristics, while the GDPR and Data Protection Act 2018 have introduced privacy restrictions which may place limitations on the use personal data for the purposes of mitigating algorithmic bias.

- **The way systems work is sometimes opaque for both technical and proprietary reasons, making scrutiny of bias more difficult.** There is a general difficulty in assessing the current impact of algorithmic bias in most domains. Limited access to the data and systems used by organisations, as well as the use of machine learning algorithms that produce opaque models, impede public scrutiny and the detection of bias.

- **The ways in which algorithmic bias is likely to be expressed, and the consequences for individuals and groups, is highly context-specific.** In some areas there may be severely detrimental consequences for relatively small numbers of individuals, while in others there may be relatively minor consequences but which are distributed across large subsections of society.

This Landscape Summary also identifies a range of gaps where research has yet to emerge. Where possible, we suggest questions and themes which might help inform future policy decisions. Some of the unanswered questions we have identified include:

- **How can we collect more empirical evidence about how algorithmic bias is being exhibited in the world today, and how it might change in the future?** While there is a growing body of literature which explains the risks associated with algorithmic bias, and the areas in which it might be creating issues, there are relatively few studies of algorithmic bias in action, which attempt to quantify the impact it is having on decision making. There is a need for more empirical studies in order to quantify the scale and severity of issues caused by algorithmic bias today, and track how this might change over time.

- **How well do existing bias mitigation techniques work, and how can they be used together to form effective mitigation strategies?** Currently, there are already many mitigation techniques emerging, as organisations direct their efforts towards addressing algorithmic bias. However, it is likely there is still considerable work to do in order to produce comprehensive solutions to all forms of algorithmic bias. There has also been little research into how effective existing mitigation techniques are in real-world contexts,

- **Which mitigation strategies are most appropriate in particular sectors and contexts?** Given that the ways in which algorithmic bias manifests are often highly context-specific, more research is needed to determine the best approaches to mitigating bias in particular areas. It is also unlikely that all forms of bias can be entirely eliminated, in which case decisions may need to be made about what kinds and degrees of bias are tolerable in certain contexts, or indeed whether algorithmic approaches should not be used because bias cannot be entirely removed.

In order to address these questions deeper collaboration between scientists, business leaders, policy makers and the public will be needed, and the complexities of this problem need to be communicated in a clear and nuanced manner. This Landscape Summary aims to contribute to this process by summarising some of the most relevant current academic literature on the subject in an accessible form.

Background

The Centre for Data Ethics and Innovation (CDEI) is an advisory body set up by the UK Government and led by an independent board of experts. It is tasked with identifying the measures we need to take to maximise the benefits of data-driven technologies for our society and economy.[1] The CDEI has a unique mandate to advise government on these issues, drawing on expertise and perspectives from across society.

In early 2019, as part of their Review of Algorithmic Bias,[2] the CDEI commissioned the Cabinet Office Open Innovation Team to engage a team of academics led by Dr Michael Rovatsos of the Edinburgh University to conduct an assessment of the current academic, policy and other literature on the subject with the aim to identify key lessons and areas where more research is needed. We would like to thank Dr Rovatsos and Dr Brent Mittelstadt, Oxford Internet Institute, and Dr Ansgar Koene, University of Nottingham for their help in writing this report. We would also like to thank Dr Rune Nyrup, Leverhulme Centre for the Future of Intelligence, and Dr Michael Veale, Alan Turing Institute, for their aid in reviewing working manuscripts and contributing their expertise to the final publication.

[1] CDEI (2019). The Centre for Data Ethics and Innovation (CDEI) 2 Year Strategy. Available at: https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2-year-strategy/centre-for-data-ethics-cdei-2-year-strategy [accessed on 08/07/19].
[2] CDEI (2019). The Centre for Data Ethics and Innovation (CDEI) 2019/20 Work Programme. Available at: https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme [accessed on 08/07/19].

Contents

1. Introduction **7**

2. Understanding algorithmic bias **10**

2.1 Algorithmic bias, discrimination and fairness 11

2.2 The legal context in the UK 13

2.3 How do algorithms work? How might algorithmic bias occur? 15

2.4 How is our use of algorithms evolving? 18

2.5 Key questions for designers, users and decision-makers 19

2.6 Evidence of algorithmic bias 19

2.7 Evidence gaps and research challenges 20

3. Mitigation of algorithmic bias **23**

3.1 Methods 23

3.2 Statistical approaches 24

3.3 Software tools 26

3.4 Discursive frameworks, self-assessment tools and learning materials 27

3.5 Documentation standards 31

3.6 Auditing 32

3.7 Technical Standards and Certification Programmes 33

4. Financial services **35**

4.1 Background 36

4.2 Use of algorithms: examples and intensity 36

4.3 Evidence of algorithmic bias in financial services 38

4.4 Case study: Algorithmic Redlining 39

4.5 Challenges and gaps 41

5. Local Government **42**

5.1 Background 42

5.2 Uses of algorithms: examples and intensity 43

5.3 Evidence of algorithmic bias in local government 44

5.4 Case Study: Child Welfare 47

5.5 Challenges and gaps 49

6. Crime and Justice **50**

6.1 Background 50

6.2 Uses of algorithms: examples and intensity 51

6.3 Evidence of algorithmic bias 53
6.4 Case study: Algorithmic risk assessments 56
6.5 Challenges and gaps 59

7. Recruitment 60
7.1 Background 61
7.2 Use of algorithms: examples and intensity 61
7.3 Evidence of algorithmic bias 62
7.4 Case study: Recruitment sifting 64
7.5 Challenges and gaps 65

8. Conclusion 66

Glossary 69

1. Introduction

Recent estimates suggest that humanity now generates around 2.5 quintillion bytes of data every day,[3] enough, if printed, to create a pile of paper that would stretch nearly one and a quarter times around the Earth.

A significant share of this information is personal data, created as organisations track and store information about our lives, including health records, banking transactions and online activity.[4] With a growing share of this personal data now being used by businesses and governments to influence and inform decisions about the general population, some commentators and academics are beginning to flag the risks. As Safiya Noble argues in her 2018 book *Algorithms of Oppression*:

> "The near-ubiquitous use of algorithmically driven software, both visible and invisible to everyday people, demands closer inspection of what values are prioritized in such automated decision-making systems."[5]

People typically have very limited awareness of the amount of data collected about them or how it can be used to feed into algorithms and shape their future interactions with a product or service. For example, a 2017 survey by The Royal Society showed that while 89% of respondents were aware of at least one of eight common applications of machine learning, only 3% claimed to know a 'great' or 'fair' amount about it.[6]

Increasingly, however, algorithmic decision-making and assisting systems are being used to classify individuals and predict behaviours based on patterns detected in the data collected about them. For example, banks will use algorithms to identify potentially fraudulent transactions based on previous patterns of fraudulent behavior, while video streaming platforms use algorithms to sort their customers into particular groups according to their preferences, with the aim of predicting the shows they will be interested in.[7]

The growing use of algorithms presents opportunities, as well as risks. As Cathy O'Neill explains in her popular 2016 book, *Weapons of Math Destruction*:

---

[3] Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read:https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#7c961a8760ba [accessed on: 13/06/19].
[4] Matsakis, L. (2019). The WIRED guide to your personal data (and who is using it): https://www.wired.com/story/wired-guide-personal-data-collection/ [accessed on: 13/06/19].
[5] Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
[6] Ipsos, M. O. R. I. (2017). Public views of machine learning. *Royal Society*.
[7] Royal Society Working Group. (2017). *Machine learning: the power and promise of computers that learn by example*. Technical report.

> "...mathematical models can sift through data to locate people who are likely to face great challenges, whether from crime, poverty, or education. It's up to society whether to use that intelligence to reject and punish them—or to reach out to them with the resources they need."[8]

A lot of effort and investment is already being put into harnessing the benefits of algorithms, with the UK keen to lead the way.[9]

In this report, we survey the latest thinking on one of the key risks: algorithmic bias. As will be shown, the debate is complex, evolving and still relatively underdeveloped. For example, one review describes the literature as "scattered",[10] while another argues that discussion has been hampered by different interpretations of what "bias" means.[11] The result is that, while many commentators express plausible fears, there is so far frustratingly little empirical evidence to back up their concerns.

The literature on algorithmic bias might not be as well developed as we would like, but any sensible debate must start by reviewing existing evidence and identifying the gaps that need to be plugged. This report does that by first explaining what algorithmic bias is, why it is causing concern, and why it is a challenging topic to research. We then look at the literature on how algorithms are being used, and what kind of impact they are having, in four key areas—finance, local government, crime and justice and recruitment. Finally, we explain the latest thinking on how the risk of algorithmic bias might be mitigated.

To narrow the focus of our work, we have largely restricted ourselves to problems of unintended algorithmic bias related to protected characteristics (i.e. where there is a legal imperative under equality law to avoid discrimination). This may not be the limit to algorithmic bias that might be considered illegal under other regimes, such as consumer law, data protection or unfair practices.[12] We nevertheless acknowledge that there will be many cases in which unethical approaches are legally permitted, and addressing algorithmic bias in the long term will require considerations beyond the letter of the law. For the most part, we leave aside other forms of bias, including where the bias is effectively intended as part of optimisation practices (e.g. a company using intentionally biased hiring practices) and areas of deliberate positive discrimination (e.g. providing specific state support to ethnic minorities).[13] More broadly, we

---

[8] O'Neil, C. (2016). *Weapons of Math Destruction*. UK: Penguin Books, p 118

[9] Great Britain. Department for Business, Energy and Industrial Strategy. (2017). Industrial Strategy: building a Britain fit for the future.

[10] Springer, A., Garcia-Gathright, J., & Cramer, H. (2018). Assessing and Addressing Algorithmic Bias-But Before We Get There... In *2018 AAAI Spring Symposium Series*.

[11] Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In *IJCAI* (pp. 4691-4697).

[12] Clifford, D. (2019). The Legal Limits to the Monetisation of Online Emotions. PhD Thesis. KU Leuven.

[13] Overdorf, R., Kulynych, B., Balsa, E., Troncoso, C., & Gürses, S. (2018). POTs: Protective Optimization Technologies. *arXiv:1806.02711 [Cs]*. Retrieved from http://arxiv.org/abs/1806.02711; Verwer, S., & Calders, T. (2013). Introducing Positive Discrimination in Predictive Models. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases* (pp. 255–270). https://doi.org/10.1007/978-3-642-30487-3_14.

acknowledge that focusing on issues of algorithmic bias can obscure challenges of discrimination and bias which result from broader systems, infrastructures, and the deployment of technology more generally. While algorithmic bias, and technical approaches to understanding it, are an important piece of tackling modern forms of discrimination and disadvantage, they are only part of the puzzle, and efforts should often reflect on the broader social and technical systems that algorithmic systems form part of.[14]

[14] Gangadharan, S. P., & Niklas, J. (2019). Decentering technology in discourse on discrimination. *Information, Communication & Society*, *22*(7), 882–899. https://doi.org/10.1080/1369118X.2019.1593484

2. Understanding algorithmic bias

Chapter summary

- Algorithmic bias can be defined in a variety of specific, technical ways, but is increasingly being used in reference to fairness and discrimination.
- There is no one generic approach to fairness, only alternative interpretations (e.g. procedural versus outcome fairness)—we cannot expect machines to reconcile these differences when society has not, and there will be trade-offs in any chosen approach.
- Legislation in the UK has certain implications for algorithmic bias—the Equality Act 2010 prohibits discrimination against people on the basis of certain protected characteristics, while the GDPR and Data Protection Act 2018 have introduced privacy restrictions which must be considered when making assessments for algorithmic bias.
- Algorithmic bias is a longstanding issue, but the growing prevalence of algorithmic decision-making systems in all aspects of life, and the increasingly complex mechanics of machine learning systems, is making it simultaneously more important and more difficult to address these issues.
- Despite the growing number of examples of algorithmic bias in daily life, there have been relatively few systematic, empirical studies of the issue—it can be difficult to access the systems and the datasets they use, and, even when that is possible, understanding what has gone wrong can be fiendishly difficult.

In 2013, Eric Schmidt, then Executive Chairman of Google, wrote that, "We have only begun to encounter the realities of a connected world: the good, the bad and the worrisome."[15] The digital age has unleashed a rapid process of change that we are struggling to understand and adapt to. Our enthusiastic adoption of new technologies is mixed with increasing concern about their side effects. And while a growing chorus of commentators have begun to express their fears, the arguments are often complicated and the evidence behind their claims can be less clear than we would like.

The debate on algorithmic bias is no different. A growing number of books, articles and reports have begun to raise concern about the risks, but the literature is complex and the empirical evidence is patchy. Nevertheless, a body of work stretching back 20 years[16] and growing quickly

[15] Schmidt, E., & Cohen, J. (2015). The new digital age: Reshaping the future of people, nations and business.
[16] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330-347.

contains many valuable questions and insights that can help us understand the nature of the problem and the gaps that still need to be filled.

2.1 Algorithmic bias, discrimination and fairness

Technically, if an algorithm produces results that are on average skewed or incorrect with respect to the population it is being used to analyse, then the conclusions are considered to be biased.[17] Colloquially, however, algorithmic bias is more commonly used to describe systematic *discrimination* on the basis of these results.

Generally speaking, discrimination can be defined as an "unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category".[18] Therefore, a reasonable definition of algorithmic bias in the sense we are using it here is the unfair treatment of a group (e.g. an ethnic minority, gender or type of worker) that can result from the use of an algorithm to support decision-making.

Discrimination and fairness are central issues here. In order to be useful, algorithms must filter or discriminate between individuals in a population (e.g. they must be able to provide a reasonable assessment of someone's credit worthiness). The central question is whether they can do this fairly.

At the most basic level, a distinction between procedural and outcome fairness is often made. Procedural fairness is concerned with the fairness of the steps, input data, and evaluations made in a decision-making process. In a data science context, this could mean an algorithm which processes data about individuals in the same way, regardless of characteristics such as gender and ethnicity. Procedural fairness also encompasses other issues, such as the input of stakeholder groups in rule-making and revision processes, and the ability for individuals to appeal decisions and easily subject them to legal scrutiny. On the other side, outcome fairness addresses the equity of the outcomes of a decision-making process, and how they are distributed across individuals and social groups within the population. It is often discussed in terms of discrimination and the denial of opportunities or services to specific groups.

A major problem is that these approaches to fairness are often fundamentally incompatible, meaning that they require judgement to choose between the most appropriate approach for a given task. For example, an employer may emphasise procedural fairness to ensure that all job applicants are treated equally, but then end up with a shortlist biased for or against certain social groups due to the use of selection criteria (e.g. education) that are proxies for membership of one group or another.

---

[17] Olhede, S., & Wolfe, P. (2018). Can algorithms ever be fair?: https://www.lawsociety.org.uk/news/blog/can-algorithms-ever-be-fair/ [accessed on: 13/06/19].
[18] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Indeed, appropriate approaches to fairness will often be highly context specific. Binns, for example, in noting the longer political and philosophical debates on fairness which data science is now drawing on, highlights one such debate on whether a single calculus can be applied to different social contexts, or whether there are "internal 'spheres of justice' in which different incommensurable logics of fairness might apply, and between which redistributions might not be appropriate".[19] One example is the difference between tests for job applications, which in principle are generally deemed fair, and tests prior to voting—while in both there is equality of opportunity, in the latter it would generally be considered that voting should not depend on talent and effort, in the same way that jobs are.

Measuring whether an algorithm is fair to different groups is challenging because a range of different statistical definitions of group fairness exist.

In particular, three formal definitions are commonly used:

1. **Anti-classification:** The model is fair if it does not use protected characteristics or proxies from which protected characteristics can be inferred.
2. **Classification or outcome error parity:** The model is fair if protected groups receive equal proportion of positive outcomes, or equal proportion of errors.
3. **Calibration:** An algorithm is well-calibrated if the risk scores it gives to people reflect the actual outcomes in real life for the people given those scores. Equal calibration definitions of fairness say that an algorithm should be equally calibrated between protected groups. For example, among those given a particular risk score, the percentage which then results in the predicted outcome should be the same between protected groups (e.g. men and women).

The problem with the use of these group measures, is they are often mutually incompatible. In particular, classification error parity (2) is incompatible with calibration (3) – an issue which is at the heart of the ongoing debate about the fairness or otherwise of the COMPAS recidivism tool. Anti-classification can also serve poorly the groups which it is designed to protect. For example, women have lower reoffending rates than men, meaning that if gender is excluded from a recidivism tool, they will overall receive disproportionately higher risk ratings.[20]

Even within types of fairness classification, tensions can exist. For example, classification error parity can involve:

---

[19] Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. Proceedings of Machine Learning Research 81:149-159, 2018 Conference on Fairness, Accountability, and Transparency: https://arxiv.org/abs/1712.03586 [accessed on: 19/06/19]; Also see Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2018). A moral framework for understanding of fair ML through economic models of equality of opportunity: https://arxiv.org/abs/1809.03400 [accessed on: 19/06/19]

[20] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. (2016). 'A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.' https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.9fa8d9982634

- **false positive error rate:** where the case is classified as positive, but in reality should have been negative.
- **false negative error rate:** the inverse—where a case is classified as negative, but should have been positive.

Ideally, a 100% accurate system would have a zero error rate, i.e. neither false positives or false negatives. In reality, this is virtually impossible to achieve, and the ratio between false positives and false negatives may vary in different cases.[21] As discussed by Kraemer et al., there is no objective fact stating whether a false negative is better than a false positive, as different users may have different preferences.[22] Therefore, the developers of these systems need to settle on a compromise where they achieve an appropriate balance between the rate of false positives and false negatives. The question of fairness is closely linked to this trade-off and will depend heavily on the specific context in which the system is used. These factors must be considered carefully before any balance is made. Within certain contexts it may be deemed fairer to optimise the system for a low rate of false negatives, or the opposite may be preferred. For an example of this trade-off in practice see the discussion of the HART system, in chapter 6.

The implication of this is that algorithms cannot be made 'fair' generically or be optimised towards all metrics of 'fairness' simultaneously. Rather, discussion needs to occur on what reasonably constitutes fairness within specific decision-making contexts. As Tene and Polonetsky conclude:

> "An ethical assessment of machine learning requires a coherent theory of discrimination. [A machine] cannot determine whether a distinction is ethical or not… We certainly should not expect the machine to make moral decisions that we have yet to make."[23]

2.2 The legal context in the UK

Regardless of questions surrounding the moral definition of fairness, there are certain key pieces of legislation in the European and UK contexts which must be observed by humans and machines alike. The first is equality legislation: in the UK this means the Equality Act 2010, which defines nine 'protected characteristics' upon which basis it is illegal to discriminate against an individual. These are:

- age
- disability
- gender reassignment

[21] Kraemer, F., Van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms?. *Ethics and Information Technology*, *13*(3), 251-260.
[22] Kraemer, F., Van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms?. *Ethics and Information Technology*, *13*(3), 251-260.
[23] Tene, O., & Polonetsky, J. (2017). Taming the Golem: Challenges of ethical algorithmic decision-making. *NCJL & Tech.*, *19*, 125.

- marriage and civil partnership
- pregnancy and maternity
- race
- religion or belief
- sex
- sexual orientation.[24]

However, it should be noted that there are certain situations where exceptions apply, regarding the use of protected characteristics. For example, in employment, religion may be considered if the job requires the person to be of a specific faith e.g. a rabbi or minister.

At first, this would appear to present a relatively straightforward framework for tackling algorithmic bias: just as humans should be held to account if they make a discriminatory decision based on a protected characteristic, so an algorithm can be programmed to explicitly exclude consideration of these attributes when arriving at a decision. In practice, though, as we will see later in the report, this is not so simple, as many of these attributes will be strongly correlated within datasets. For example, while an algorithm may be explicitly programmed to disregard data on ethnicity, if this data is strongly correlated with another attribute, such as postcode, then the algorithm may still come to racially-biased decisions, but by proxy rather than directly. Additionally, protected features may often need to be included within datasets, in order to provide adequate measures of bias and mitigation within the algorithmic system.

The EU General Data Protection Regulation, and the Data Protection Act 2018 (which implements the provisions of the GDPR, the Law Enforcement Directive and the Council of Europe's Convention 108+ in UK law, which in this area is expected to remain broadly consistent for the foreseeable future, regardless of Brexit) carry with them significant implications for algorithmic bias. At least in theory, these regulations mandate a "right to explanation" with regards to automated decision-making, and the right to opt out of automated decisions.[25] They also build on provisions in the Data Protection Act 1998 to implement stricter rules requiring that *explicit* consent be obtained from individuals for the purposes to which certain 'sensitive' personal information will be put, which can carry significant implications for data scientists attempting to check systems for bias.

Sensitive data is often required for understanding and seeking to mitigate algorithmic bias, and this can pose practical challenges as well as legal ones.[26] On first glance, it would seem that checking for algorithmic bias using data revealing ethnicity or health, for example, would require

---

[24] Equality and Human Rights Commision (2019). Protected Characteristics: https://www.equalityhumanrights.com/en/equality-act/protected-characteristics [accessed on: 13/06/19].
[25] Although scholars have questioned whether, due to the wording of these particular clauses, this will actually be meaningful or enforceable in practice. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76-99.
[26] Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 205395171774353. https://doi.org/10/gdcfnz

explicit consent of the data subject under the Data Protection Act 2018 and the GDPR.[27] Yet the restrictions in the Data Protection Act 2018 on processing sensitive data do come with exemptions for processing which is necessary for the purposes of identifying or keeping under review the existence or absence of equality of opportunity or treatment between groups of people specified in relation to that category with a view to enabling such equality to be promoted or maintained.'[28] Further exemptions can be made by the Secretary of State by regulations if deemed required in the area of debiasing without changing the structure of data protection legislation.

In the public sector, the 'public sector equality duty' in the Equality Act 2010 requires public authorities to have due regard to issues of discrimination while exercising all their functions, and is likely to require careful consideration of algorithmic bias concening protected characteristics within the procurement, deployment and maintenance of algorithmic systems.[29] Scotland has commenced the 'socioeconomic equality duty' in addition, which requires a consideration of socioeconomic disadvantage in equality law in a similar way.[30] Both legal obligations seem likely to require a strong awareness of issues of algorithmic bias in order to carry out fully and with rigour.

2.3 How do algorithms work? How might algorithmic bias occur?

Algorithms are processes to be followed in a problem solving operation or calculation. Machine learning algorithms are mathematical models designed by humans to draw conclusions or make predictions by analysing data considered relevant to a question under consideration (e.g. who should we shortlist for this job?). Put another way, these models are "opinions embedded in mathematics", and can be as subjective as the assumptions that went into their creation.[31] At its simplest level, data is fed into a model, mathematical processes are applied, and a corresponding output is created (Figure 1).

[27] GDPR, art 9.
[28] Data Protection Act 2018 sch 1 para 8(1)(b).
[29] Law Society Commission on the Use of Algorithms in the Justice System. (2019). *Algorithms in the Criminal Justice System*. London: The Law Society of England and Wales. London, pp. 30, 68-70.
[30] Equality Act 2010, s 1; ibid p. 70.
[31] O'Neil, C. (2016). *Weapons of Math Destruction*. UK: Penguin Books.

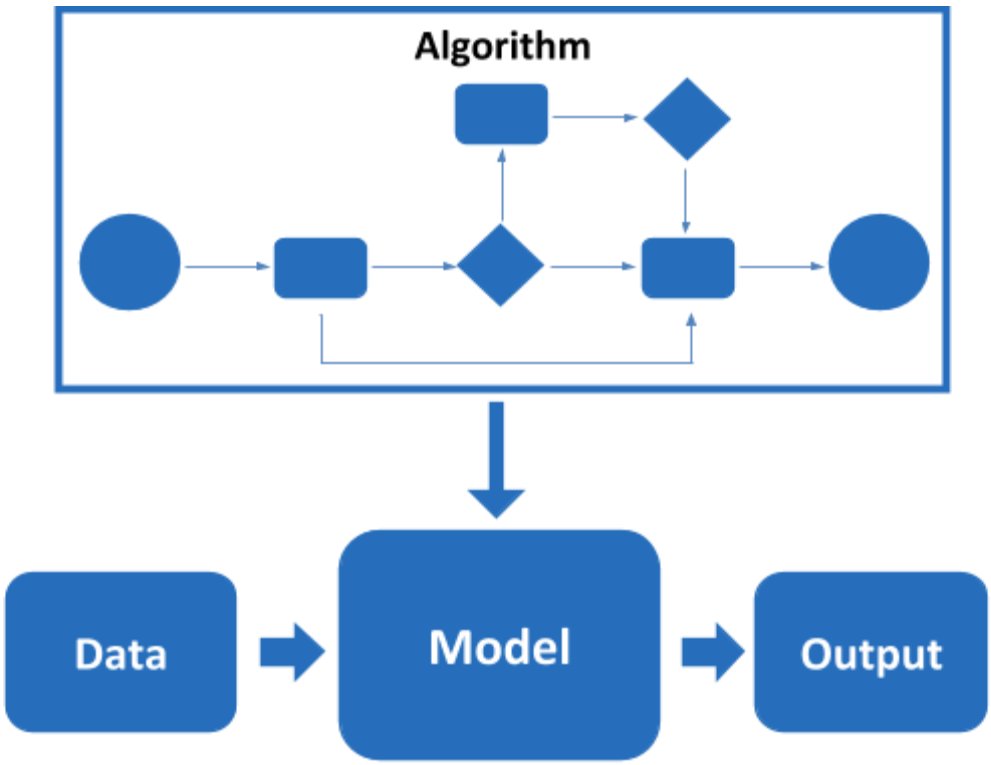

Figure 1: How an algorithmic model works

In particular, the development of machine learning algorithms typically goes through several stages, from input to the utilisation of the outputs in a repeating loop as new data is fed back into the algorithm with the aim of improving its accuracy (see Figure 2). As David Leslie notes:

> "Human error, prejudice, and misjudgement can enter into the innovation lifecycle and create biases at any point in the project delivery process from the preliminary stages of data extraction, collection, and pre-processing to the critical phases of problem formulation, model building, and implementation."[32]

[32] Leslie, D. (2019). Understanding artificial intelligence ethics and safety. *Alan Turing Institute,* 13-22. Also see Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In *IJCAI* (pp. 4691-4697); Silva, S., & Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon (1960-)*, *55*(1 & 2), 9-37.

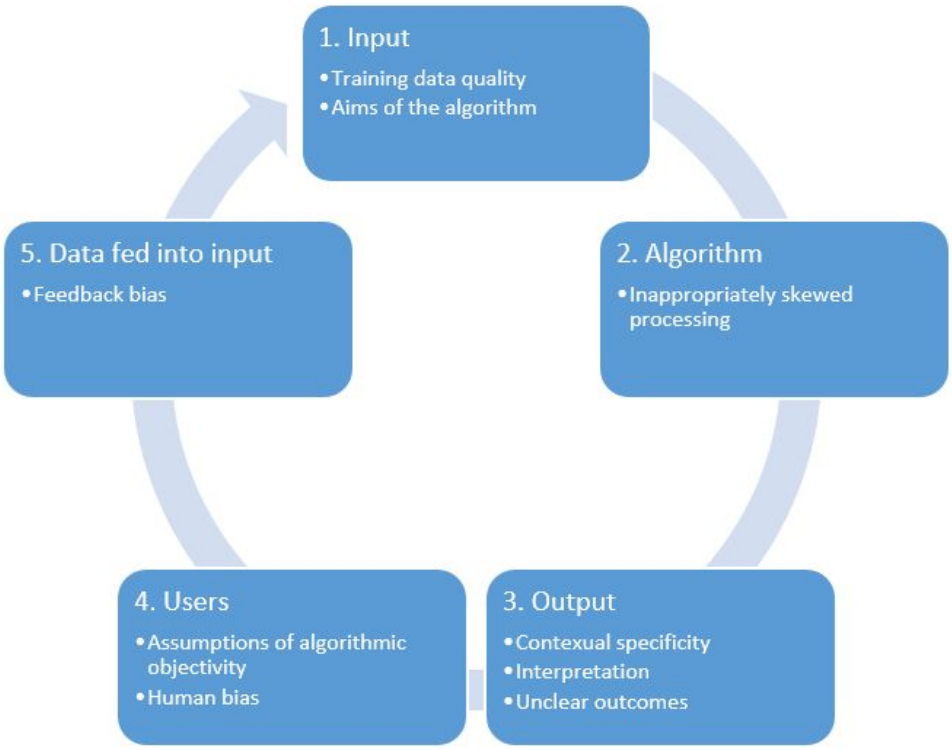

Figure 2: The algorithmic process and the potential biases introduced at each stage (credit to Danks & London (2017) and Silva & Kenney (2018)).

During the input stage (1), the algorithm is trained on data that will inform the resulting calculations, outputs and decisions.[33] Therefore, if the original training data is biased, the algorithm will perpetuate and potentially compound these biases. As John Giannandrea, Apple's Senior Vice President for Machine Learning and AI Strategy, has stated in the past, "The real safety question, if you want to call it that, is that if we give these systems biased data, they will be biased."[34]

Bias can also be introduced in the design of an algorithm (2); for example, when variables are weighted incorrectly by programmers, whether intentionally or unintentionally.[35] In fact, for Mittelstadt et al "algorithms are inescapably value laden" with "operational parameters… specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others."[36]

[33] Eaglin, J. M. (2017). Constructing recidivism risk. *Emory LJ*, *67*, 59.
[34] Knight, W. (2017). Forget Killer Robots—Bias Is the Real AI Danger: https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/ [accessed on: 13/06/19].
[35] Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In *IJCAI* (pp. 4691-4697).
[36] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

The use of an algorithm's output can also be subject to bias (3 and 4) resulting from how humans interact with the outputs and the false assumption that using algorithms produces neutral results.[37] For example, a probation officer who is assisted in her work by an algorithm which produces risk ratings for offenders may, without a proper understanding of the system, come to believe that the algorithm is entirely objective and infallible, and automatically accept suggestions made by the system without considering alternative assessments which she could draw from her own knowledge and experience.

Many systems are also further trained and optimised over time, usually based on data concerning whether the original recommendations or predictions were judged to be accurate or useful by human operators (5). This means that any biases that were generated in the previous four steps can be further amplified over time.

2.4 How is our use of algorithms evolving?

The growth of machine learning (ML) is increasing the risk of algorithmic bias. While conventional algorithms can be incredibly complex, they still consist of a sequence of steps that have been manually coded by human programmers, who should understand the purpose and logic behind every step.[38]

By contrast, ML algorithms are given an objective and 'training' data sets, and then 'learn' from this to write their own sequence of steps which will get them to the desired outputs—essentially a process of automated reverse engineering. As such, the underlying process used by the machine once it has been trained is often not well understood, and this can make figuring out potential sources of bias, and where exactly it has entered the system, much more complicated. This problem is compounded in the case of deep learning algorithms, which use many layers of computational process to reach their end result, with the purpose behind individual step or calculation potentially meaningless to a human observer.[39]

Furthermore, the context in which we are using these algorithms is starting to change as well. Increasingly we are relying on algorithms to either make or support operational decisions. This report refers broadly to algorithmic decision-making systems, which we use to refer to a range of tools, ranging from systems which might provide advice to a human decision-making (as with many risk assessment tools), through to systems which might make a decision with almost no human input or oversight (as with systems which automate decisions about consumer finance, for example). The distinction between these can be difficult to pin down—automation bias, whereby human users who are provided with advice by machines will often become increasingly

---

[37] Bogen, M., & Rieke, A. (2018). Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias: https://apo.org.au/sites/default/files/resource-files/2018/12/apo-nid210071-1229641.pdf [accessed on: 13/06/19].

[38] Bruckner, M. A. (2018). The promise and perils of algorithmic lenders' use of big data. *Chi.-Kent L. Rev.*, *93*, 3.

[39] Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale JL & Tech.*, *20*, 103.

reliant on and uncritical of this advice with time, means a decision-assisting tool can easily become a decision-making tool in practice. While this review focuses mostly on biases embedded within algorithmic systems and their data, it is important to be aware of biases which can be introduced in the interactions between humans and algorithmic systems as well, which we consider where relevant.

2.5 Key questions for designers, users and decision-makers

The risk of algorithmic bias, and the potential for it to grow significantly as the adoption of algorithmic decision-making tools increases, raises a number of challenging questions for policy makers, businesses, academics and citizens. These include:

- **How can algorithmic biases be detected and measured?** Are algorithmic biases fundamentally different or worse when compared with pre-existing human biases? What tests and measurements are needed, and what scrutiny of data and algorithms is needed? How do we define unacceptable levels of bias in a system?

- **How can algorithmic biases be mitigated and/or regulated?** Should new or existing regulators be empowered to enforce measures to detect and mitigate biases? Where algorithmic biases cannot be mitigated or accounted for, should limits be placed on the use of these systems? Does the increased use of algorithms mandate the creation of new legislation, or can existing legislation be amended, to account for regulatory gaps?

- **What are the implications for social justice and responsibility?** If there are shared societal risks emanating from algorithmic bias, what is the appropriate balance of liability between different stakeholders? Who bears the responsibility for ensuring compliance of algorithmically supported processes with existing policy and law?

2.6 Evidence of algorithmic bias

Any reasonable debate on such an important set of questions should be informed by the best available evidence. Algorithmic bias is an area of growing interest for academics, think tanks, and journalists, whose combined efforts are beginning to shed light on the issue. For example:

- A 2015 study showed that many of the algorithms used by insurance companies in the US to create quotes for car insurance were relying on credit scores more heavily than driving records. This meant that in Florida, an individual with a clean driving record but poor credit score could end up paying $1,552 more for car insurance than the same driver with a drink driving conviction but an excellent credit score.[40]

- Researchers at MIT found that three of the latest gender-recognition programmes, developed by IBM, Microsoft, and Megvii, could correctly identify a person's gender from

[40] O'Neill, C. (2016). *Weapons of Math Destruction*. UK: Penguin Books, pp 164-165.

a photograph 99% of the time – but only for white men. For BAME women, accuracy dropped to just 35%.[41]

- In 2016 it was reported that Amazon Prime's same day delivery services were much less likely to be offered to customers in predominantly Black and Hispanic neighbourhoods in a selection of US cities, including New York, Atlanta and Boston.[42] Amazon has since begun delivering to many of these neighbourhoods and said "that its [initial] delivery decisions [weren't] based on the ethnic composition of a neighborhood, but several factors including the concentration of Prime members, as well as the proximity of the area to Amazon's warehouses."[43] See Chapter 3 for more on this.

- A 2015 paper by Ammit Datta and others found that use of Google's Ad Settings feature can lead to "seemingly discriminatory ads." For example, the authors found that visiting web-pages associated with substance abuse changed the ads shown and that setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male.[44]

2.7 Evidence gaps and research challenges

Despite the growing interest in algorithmic bias, much of the literature is speculative and rarely based on the kind of concrete evidence described above. One reason for the evidence gap is that academic research has been limited by an inability to access and systematically audit the algorithms used by proprietary tools.[45] Many companies are highly protective of the algorithms they use, especially where their market success is dependent on them and could be placed in jeopardy if a competitor is able to replicate them. For example, the exact composition of Google's PageRank algorithm, which attempts to display websites in order of their importance in search results, is a closely guarded secret.[46] For an external researcher, exposing potential biases it may have with any precision is therefore a challenging task.

In the case of deep learning algorithms, it can be extremely difficult even for the companies in question to understand exactly how their algorithms have come to their outputs. Google uses

---

[41] Revell, T. (2018). Face-recognition software is perfect – if you're a white man: https://www.newscientist.com/article/2161028-face-recognition-software-is-perfect-if-youre-a-white-man/ [accessed on: 13/06/19].
[42] Ingold, D., & Soper, S. (2016). Amazon doesn't consider the race of its customers. Should it?: https://www.bloomberg.com/graphics/2016-amazon-same-day/?cmpid=google [accessed on: 13/06/19].
[43] Banchiri, B. (2016). Is Amazon same-day delivery service racist?: https://www.csmonitor.com/Business/2016/0423/Is-Amazon-same-day-delivery-service-racist [accessed on: 13/06/19].
[44] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, *2015*(1), 92-112.
[45] Springer, A., Garcia-Gathright, J., & Cramer, H. (2018, March). Assessing and Addressing Algorithmic Bias-But Before We Get There... In *2018 AAAI Spring Symposium Series*.
[46] DeMers, J. (2018). How Much Do We Really Know About Google's Ranking Algorithm?: https://www.forbes.com/sites/jaysondemers/2018/02/07/how-much-do-we-really-know-about-googles-ranking-algorithm/#30fa9fc955bb [accessed on: 14/06/19].

deep learning algorithms to identify images, but each identification is based on many smaller decisions which, taken in isolation, would seem meaningless to a human being. Even Google's own developers appear to struggle with the task. In 2015, a software developer highlighted the fact that Google's visual identification algorithm could not accurately distinguish between Black people and gorillas. Three years later, it was found that Google had simply switched off the ability to search for gorillas and associated terms in products such as Google Photos which use this feature, rather than fix the algorithm.[47]

Other barriers to understanding algorithmic bias include:

- **Disparate terminology:** a common vocabulary for discussing these issues is only just beginning to emerge. This is particularly pertinent for the issue of algorithmic bias, which affects many different technical systems and is discussed across a variety of academic literatures. At the moment, it is possible to find discussions under headings such as algorithmic bias, algorithmic fairness, big data, data transparency, AI ethics and many others.

- **Difficulties in accessing the data on which an algorithm has been trained:** in many cases, companies may be as protective of their datasets as they are of their algorithms, as these may have taken years to accumulate and provide a significant market advantage.

- **The quality of historical data, and a limited understanding of biases which existed at the time of collection:** the way that historic data has been collected by governments or police authorities, for example, may have been significantly influenced by official or unofficial policies or even individual decisions made at the time of collection, which could be difficult or impossible to detect after the fact.

- **Difficulties in anticipating the future behaviour of an algorithm without knowing what further data will be used for its training:** algorithms will often be repurposed after their initial development, which may involve them being trained on an entirely new dataset, with its own potential biases. A system can be tested with hypothetical data to test potential future situations and purposes, but this is unlikely to ever be exhaustive.

Research on algorithmic bias is challenging, but the debate has been more intense in some areas than in others. After discussing ways to mitigate bias, the remaining four chapters deal in turn with algorithmic bias in financial services, local government, crime and justice and recruitment.

Although some of the issues discussed may appear in more than one of these sectors, it is important to highlight that they will often have very different implications for society, including the portion of the population that are affected and the level of impact that these issues have in

[47] Authority of the House of Lords (2018). AI in the UK: ready, willing and able?: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf [accessed on: 13/06/19].

specific contexts. For example, the impact of bias in crime and justice may be higher for an individual than in recruitment. Receiving a harsher prison sentence, due to bias, would arguably have a much greater effect on a person's life than being rejected for a specific job application. In addition, in one area the impact may be substantial for a small number of people, while in another the impact could be low but affect a large group. This distinction is not trivial, and it is often not immediately obvious what would constitute the 'fairest' or most desirable outcome. Therefore, these factors need to be examined carefully within each specific context, in order to design an appropriate response.

3. Mitigation of algorithmic bias

Chapter summary

- There are a range of approaches to mitigating algorithmic bias, and each have their own advantages and limitations. However, there is little current research into how mitigation strategies work in practice, or within particular domains, where the appropriateness of particular techniques is likely to vary considerably.
- The first commercial and open source software tools for algorithmic bias analysis are appearing, but are still relatively untested. Organisations have also developed approaches and processes for helping software developers think about possible sources of bias, and engage with stakeholders who may be affected by them. Documentation about datasets can help software developers understand where data has come from and its possible issues.
- Public and private organisations are developing various technical standards and certification programmes around algorithmic bias, but more work needs to be done to ensure they align with one another, and that each has a clear and distinct purpose.
- Access to personal data concerning protected characteristics is needed for some approaches to mitigating bias, but this remains a major challenge for data protection reasons.

The development and adoption of mechanisms to measure and detect fairness and bias is at an early but encouraging stage. Legal and ethical governance of algorithmic systems is increasingly an international policy priority—over 18 countries are currently developing national AI strategies, and almost all include some provision for the development of ethical standards or approaches.[48] Growing commercial and professional interest is demonstrated by the involvement of major professional bodies and technology companies in initiatives such as the 'Partnership on AI to benefit people and society'.[49]

3.1 Methods

Methods to detect and mitigate biased, unfair, and discriminatory decision-making broadly fall into five categories:

[48] Notable exceptions include South Korea and Taiwan. Dutton, T., Barron, B., & Boskovic, G. (2018). Building an AI World: Report on National and Regional AI Strategies.
[49] Partnership on AI: https://www.partnershiponai.org/ [accessed on: 14/06/19].

- statistical approaches and software toolkits;
- discursive frameworks, self-assessment tools and learning materials;
- documentation standards;
- auditing;
- the development of technical standards and certification.

These are not mutually exclusive, and particular initiatives often contain several of these elements. They each have their own respective strengths and weaknesses, and there is no one technique which is likely to address all potential issues of algorithmic bias. One of the key distinctions is between broadly technical approaches, such as the use of statistical and software-based techniques, and discursive strategies.

Technical approaches can in theory offer a more consistent and efficient approach, but may well struggle in situations where there is no clear definition of what constitutes fairness—given the precise instructions they must work to, an algorithmic system cannot navigate moral grey areas in the same way that humans can. Conversely, while discursive strategies, such as workshops and discussion forums, and ensuring humans are overseeing systems and retain the ability to override automated decisions where necessary, may be costly and inconsistent, they can also deal with situations in which machines would struggle, as humans generally have an intuitive sense of what is fair and what is not (often in the sense that they 'know it when they see it'). What constitutes an appropriate bias mitigation strategy is therefore likely to vary considerably based on the context. A strategy which is acceptable for mitigating bias in the context of allocating drivers in a ride-sharing app is likely to be very different to the steps which should be taken to mitigate bias in risk profiling systems used in the criminal justice system, for example.

Finally, there are also those who question the techno-centric framing of algorithmic bias more fundamentally. This school of thought argues that by privelaging tech-mediated forms of bias as distinctive from, and perhaps more important than, conventional or unmediated forms of bias, discrimination and predjudice in the world today, there is a risk that we become overly focused on narrowly construed technical fixes.[50] This, it is argued, comes at the expense of a wider examination of the sources of discrimination and injustice, which while further perpetuated by technology are not in themselves generated by it. They call for a shift in focus away from tech-mediated discrimination, to a consideration of algorithmic bias as just one facet in a wider ecosystem of injustice.

[50] Peña Gangadharan, S., & Niklas, J. (2019). Decentering technology in discourse on discrimination. *Information, Communication & Society*, *22*(7), 882-899.

3.2 Statistical approaches

Recent years have seen the development of a multitude of statistical approaches and bespoke algorithms to detect and prevent bias and unfair or discriminatory decision-making.[51] Broadly speaking, these methods either focus on identifying patterns of discrimination in historical datasets before it is used to train an algorithm, or by adjusting models and their outputs to be non-discriminatory.[52] The three strategies below reflect the implementation of such methods at different stages of data analysis and decision-making:[53]

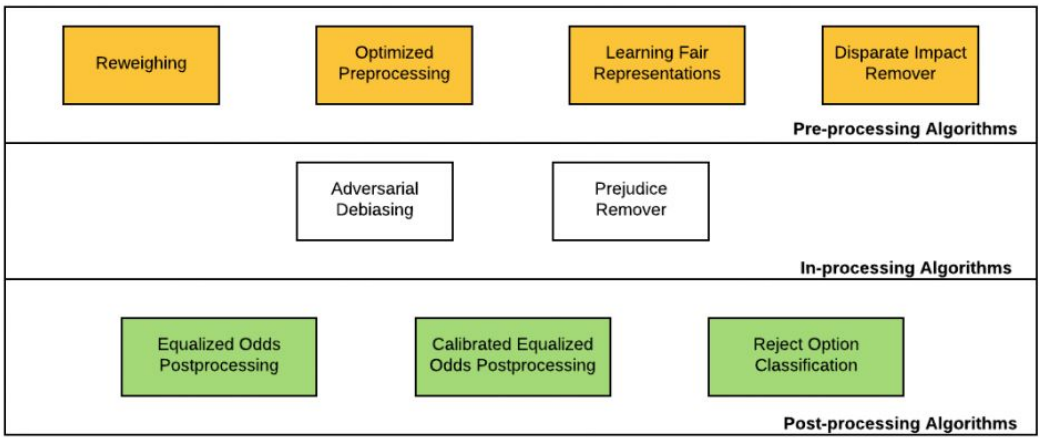

Figure 3: Methods of mitigating bias with ML models (the above figure was taken in its entirety from the article by Ajitesh Kumar (2018).[54]

Pre-processing methods involve modifying the training data, with the aim of preventing the algorithmic model from learning discriminatory decision-making rules in the training stage. This can be accomplished by, for example, modifying the training data itself,[55] for instance by

[51] Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, December). Considerations on fairness-aware data mining. In *2012 IEEE 12th International Conference on Data Mining Workshops* (pp. 378-385). IEEE.

[52] For those interested in the technical details of such methods: Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582-638, offer a comprehensive review of discrimination analysis in data collection and analysis. Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, offers a comprehensive survey of methods for measuring and detecting indirect discrimination in machine learning. Custers, B. H. M., Calders, T., Schermer, B. W., & Zarsky, T. Z. (1866). Discrimination and privacy in the information society. *Studies in applied philosophy, epistemology and rational ethics*, *3*, provides an in-depth overview of methods of discriminatory prevention and discovery.

[53] Kumar, A. (2018). Machine Learning Models: Bias Mitigation Strategies: https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies [accessed on: 14/06/19].

[54] Kumar, A. (2018). Machine Learning Models: Bias Mitigation Strategies: https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies [accessed on: 14/06/19].

[55] d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, *5*(2), 120-134.

changing the values of specific attributes for individual records[56] or even removing attributes entirely.

In-processing methods involve modifying the algorithmic model itself. One approach is to train separate models for each protected group in isolation of one another, and then only use the relevant model for decisions concerning that group—this would, however, be difficult in many situations where certain individuals may belong to multiple categories (Asian and female, for example).[57] Another is to change the criteria that result in 'branches' in a decision tree in order to ignore or correct the influence of protected characteristics.[58]

However, many in-processing methods require personal data regarding protected characteristics to be available, which cannot be taken for granted due to the legal sensitivity of this data. The legal status and necessity of monitoring for bias means well-intentioned data scientists wishing to detect discrimination in their systems can face barriers to obtaining the necessary data or information about protected characteristics, especially if they did not seek consent to gather and process the data for these purposes from the beginning.[59]

Post-processing methods involve removing discriminatory rules or otherwise modifying a model (e.g. confidence intervals, weights, probabilities, predicted classes or labels) after it has been trained. This might mean, for example, modifying a model so that it places less significance on particular postcodes, which could be closely correlated with one specific ethnic group. Outcomes or decisions can also be artificially adjusted to ensure equitable treatment across groups within the affected population. For example, if it is known that a probation risk assessment algorithm consistently ranks one ethnic group as a higher risk than others, any risk assessment relating to an individual from that group might be downgraded by a human probation officer to ensure an equitable outcome.[60]

3.3 Software tools

Software toolkits are now being developed which encompass these statistical methods for measuring and mitigating bias in algorithmic decision-making systems. While it is difficult to determine how rapidly these new toolkits are being adopted in practice, their rapid emergence suggests a significant demand in the public and private sectors. Two specific examples are

[56] Luong, L. P., & Ha, D. T. (2011). Behavioral Factors Influencing Individual Investors' decision-making and Performance. *Survey of the Ho Chi Minh Stock Exchange, Umea School of Business Spring semester*.
[57] Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2), 277-292.
[58] Kamiran, F., & Calders, T. (2010, May). Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands* (pp. 1-6).
[59] Kamiran, F., & Calders, T. (2010, May). Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands* (pp. 1-6).
[60] d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, *5*(2), 120-134.

Accenture's 'Fairness Tool', and IBM's 'AI Fairness 360 Open Source Toolkit', but further efforts can be found in Annex A.

The 'Fairness Tool' aims to identify bias and possible proxies for protected characteristics within the datasets used by algorithmic systems.[61] The tool can remove relationships between sensitive variables and proxies that can result in unfair or discriminatory outcomes. The tool can similarly balance the rates of false positives and negatives across groups. As the exclusion of particular attributes can also be detrimental to the accuracy of the wider system, it will also show the user what the potential impact of excluding characteristics will be, leaving it to the user to decide whether they want to prioritise fairness or accuracy. However, unlike several of the other toolkits and methods which are available, the 'Fairness Tool' is not freely available.

IBM has developed the 'AI Fairness 360 Open Source Toolkit' which aims to assist developers to "examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle".[62] The toolkit provides tests and algorithms to measure fairness and mitigate bias in datasets and models. In total, the toolkit allows for measurement of fairness on over 70 metrics and includes ten bias mitigation algorithms.

These toolkits are a step in the right direction, and demonstrate a commitment by researchers, industry and government bodies to take algorithmic bias and fairness seriously. However, while the use of these tools in the development stage should be encouraged, it is not enough to detect and mitigate bias post-release; algorithms 'in the wild' can develop new biases and make unfair or discriminatory decisions as they are exposed to new data and used in different decision-making contexts. Toolkits must therefore be understood as one promising avenue for mitigating bias that will only be effective if they are embedded in a long-term, iterative governance process which encompasses an algorithm's entire lifecycle.

3.4 Discursive frameworks, self-assessment tools and learning materials

Less technical methods have also emerged to measure and mitigate bias. As suggested above, lifelong and inclusive governance is required to effectively address algorithmic bias, fairness, and discrimination. A number of tools have been developed based on self-assessment, education, and interaction with stakeholders affected by algorithmic systems. Several NGOs, think tanks, government organisations and others have developed a range of such tools to identify and mitigate bias in algorithmic decision-making systems.

For example, the US-based AI Now Institute has proposed 'Algorithmic Impact Assessments (AIAs)', a self-assessment framework for public agencies to assess the potential impact of automated systems prior to procurement, similar to existing environmental and data protection

[61] Chowdhury, R. (2018). Tackling the challenge of ethics in AI: https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai [accessed on: 14/06/19].
[62] Varshney, K. R. (2018). Introducing AI fairness 360: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360 [accessed on: 14/06/19].

impact assessments.[63] The framework identifies five key elements to be undertaken by public agencies (see Box 1).

Box 1: Key Elements of AI Now's Algorithmic Impact Assessment

1. Agencies should conduct a self-assessment of existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias or other concerns across affected communities;
2. Agencies should develop meaningful external researcher review processes to discover, measure or track impacts over time;
3. Agencies should provide notice to the public, disclosing their definition of "automated decision system," existing and proposed systems, and any related self-assessments and researcher review processes before the system has been acquired;
4. Agencies should solicit public comments to clarify concerns and answer outstanding questions;
5. Governments should provide enhanced due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased or otherwise harmful system uses that agencies have failed to mitigate or correct.

AI Now subsequently released an 'Algorithmic Accountability Policy Toolkit' to aid public agencies in undertaking an AIA.[64] The Canadian government, with support from AI NOW, has adopted an AIA approach for their procurement of AI systems.[65]

In the United Kingdom, data protection impact assessments will be obligatory for most uses of machine learning involving personal data.[66] In the public sector, these documents are not public by default, but might be subject to Freedom of Information law.[67] Data protection impact assessments are not limited to data protection issues, but must consider 'risks to the rights and freedoms of natural persons' more generally. The Information Commissioner's Office highlights that the risks include 'risks to privacy and data protection rights, but also effects on other

[63] Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: a practical framework for public agency accountability. *AI Now*.
[64] AI Now (2018). Algorithmic Accountability Policy Toolkit: https://ainowinstitute.org/aap-toolkit.pdf [accessed on: 14/06/19].
[65] Government of Canada. Algorithmic Impact Assessment (Archived): https://canada-ca.github.io/digital-playbook-guide-numerique/views-vues/automated-decision-automatise/en/algorithmic-impact-assessment.html
[66] Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review*, *16*(1), 18–84. https://doi.org/10/gdxthj.
[67] Binns, R. (2017). Data protection impact assessments: A meta-regulatory approach. *International Data Privacy Law*, *7*(1), 22–35. https://doi.org/10/cvct.

fundamental rights and interests', including the 'impact on society as a whole'.[68] As a consequence, issues like discrimination must be considered as relevant, in the context of the way data protection legislation is not an instrument to protect the fundamental rights to private life and to data protection alone, but an instrument to safeguard an array of rights and freedoms in a digital age.[69]

More prospectively in the UK, David Leslie at the Alan Turing Institute has proposed Stakeholder Impact Assessments (SIAs), which aim to highlight the possible social and ethical impacts of the AI systems. Leslie also suggests that the designers of algorithmic system prepare a Fairness Position Statement (FPS) in which the fairness criteria being employed in the AI system is made explicit and explained in plain and non-technical language. This FPS is then required to be made publicly available for review by all affected stakeholders.[70]

The UnBias project, funded by the EPSRC, developed the 'Fairness Toolkit'[71] (not to be confused with Accenture's Fairness Tool) containing several tools intended to raise awareness and facilitate dialogue around algorithmic bias and fairness. Along with an explanation of the toolkit and a framework for participants to assess its utility, the toolkit contains three components:

1. **'Awareness Cards'**, which provide examples of how bias and unfairness occur in algorithmic systems. These can be used in discussions, debates and workshops.

2. **'TrustScape'**, a poster designed for the general public which provides a canvas for participants to explore their perceptions of algorithmic bias, data protection, and online safety.

3. '**MetaMap'**, a poster designed for policy-makers, regulators, members of the public sector, researchers and industry to respond to the visualisations created by participants in the TrustScape.

[68] Information Commissioner's Office (2019) What is a DPIA?. *Guide to the GDPR*. Retrieved July 5 2019.
[69] See to that effect the Court of Appeal in Dawson-Damer & Ors v Taylor Wessing LLP EWCA Civ 74 (2017), para 107.
[70] Leslie, D. (2019). Understanding artificial intelligence ethics and safety. *Alan Turing Institute*
[71] UnBias. Fairness Toolkit: https://unbias.wp.horizon.ac.uk/fairness-toolkit [accessed on: 14/06/19].

Figure 4: Examples of UnBias Project 'Awareness Cards' (reverse side shown)[72]

The Center for Democracy and Technology has also created a 'Digital Decision Tool'. This consists of an interactive flowchart designed to raise concerns regarding bias, fairness, and related ethical issues during the development and deployment phase of an algorithmic system.[73] For example, when a user considers the design of their system, and what data they will be using, they are asked to think about possible constraints, with questions such as: "are there any fields which should be eliminated from your data?".

Academics from Harvard Law School and the Berkman Klein Center for Internet and Society have also published a number of relevant documents in this area.[74] These resources look at the possible impact of AI on human rights and also compares the different principles, guidelines and frameworks that have been proposed relating to the development and implementation of AI

[72] See https://unbias.wp.horizon.ac.uk/fairness-toolkit/
[73] Duarte, N. (2017). Digital Decisions Tool: https://cdt.org/blog/digital-decisions-tool [accessed on: 14/06/19]; Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.
[74] Harvard University (2018). Artificial Intelligence and Human Rights, available at: https://ai-hr.cyber.harvard.edu/ [accessed on: 10/07/19].

technologies in recent years. These could help in identifying the emergence of sectoral norms over time.

Self-assessment, education and interactive tools will only be effective as long as they are given serious consideration throughout the lifecycle of an algorithm. Pre-deployment impact assessments can be effective at identifying types of bias and unfair decision-making in particular contexts, which may have unique historical underpinnings that are not self-evident to the software developers. However, any form of self-assessment can prove ineffective if it is not sufficiently critical and inclusive. Self-assessments and interactive approaches should be implemented with a clear route towards influencing the design and/or deployment of the algorithmic system in question, and where possible should involve external stakeholders, researchers and third-party regulators. Otherwise there is a risk that these approaches become checklists without any meaningful impact on the development and deployment of an algorithm.

3.5 Documentation standards

The growth in big data analytics and machine learning is being facilitated by the usage, sharing and aggregation of diverse datasets, sometimes with only a limited understanding of how this data has been generated and what its strengths and weaknesses are. This runs the risk that unintended biases will be introduced to the datasets or the way in which they are processed by, for example, unintentionally using a dataset without understanding its original context. As a result, efforts are now being made to standardise the documentation which comes with datasets.[75]

Documentation standards are intended to establish the basic information to be filled out when collecting new data or training a new model, which can then inform the decision-making of other developers and researchers in the future. Such information could include the creation, contents, intended uses, and any relevant ethical and legal concerns about the data. This information would help users interrogate datasets and identify potential biases in datasets and models prior to and during processing. The standardisation of documentation could also drive better data collection practices in the first place, as well as consideration of contextual and methodological biases more generally. However, at present, there is no standardised form for what information should be included about datasets.

Major companies and research bodies are starting projects to establish documentation standards. Microsoft Research and Google have both been involved in projects to draw up a standardised set of documentation, or 'datasheets', for datasets used to train algorithmic models. These documents are intended to help potential users decide "how appropriate the corresponding dataset is for a task, what its strengths and limitations are, and how it fits into the broader ecosystem."[76] These initiatives tend to focus on documenting potential biases,

[75] He, Q. (2016). Three lessons of data standardisation: https://www.linkedin.com/pulse/three-lessons-data-standardization-qi-he [accessed on: 14/06/19].
[76] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*; Holland, S., Hosny, A., Newman, S.,

gaps, proxies, and correlations which could be inadvertently picked up and reinforced by machine learning systems making use of the data.

To complement these 'datasheets', a team at Google has proposed 'model cards for model reporting',[77] a short set of documentation accompanying trained models that describes various performance characteristics and intended contexts of use. The cards evaluate how performance varies by context, for instance when applied to different cultural, demographic, phenotypic and intersectional groups. In a UK context, David Leslie has proposed that developers adopt a 'Process-Based Governance Framework', which would involve drawing up a Dataset Factsheet, Stakeholder Impact Assessment, Discriminatory Non-Harm Self-Assessment, Fairness Position Statement, Safety Self-Assessment, and a Model Sheet for Implementers for each project. All of these records would then be included in a Process Log for the system.[78]

A 'standardised declaration of conformity' for AI has been proposed to address the purpose, performance, safety, security, and provenance of AI products.[79] In contrast to other proposals, this declaration is explicitly consumer-facing in order to enhance transparency and trust with those using AI products.

3.6 Auditing

Systems can be tested for algorithmic bias through audit-based methods, in which the decisions from a system are compared to see what the outcomes are for particular groups within an affected population, based on different datasets or interactions. In practice, these tend to resemble social scientific audit studies, and can involve surveys, A/B testing, non-invasive data scraping, and crowdsourced audits in which users collect data by interacting with the system in question.[80] In such approaches, human subjects or bots mimicking human behaviour are used to collect information about the performance of an algorithmic system in terms of bias and fairness of outcomes.

However, these methods are primarily observational, insofar as they cannot tell us how a system was created, what values or requirements informed its design, or indeed explain the behaviour of the system (e.g. why a particular group was shown lower prices than another

---

Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

[77] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM.

[78] Leslie, D. (2019). Understanding artificial intelligence ethics and safety. *Alan Turing Institute*

[79] Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., & Varshney, K. R. (2018). Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv preprint arXiv:1808.07261.

[80] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, *22*.

group).[81] This is a problem, as determining the source of bias in algorithmic systems is difficult without also knowing about the system's development history and training.[82]

Additionally, biases, inaccuracies or inequitable outcomes may only emerge in relation to particular data, meaning audit studies are unlikely to ever comprehensively verify that a system is unbiased in general, but rather only within the particular context studied. This is comparable to the limitations facing pre-deployment assessments and testing for fairness as described above.

3.7 Technical Standards and Certification Programmes

A number of national and international organisations have also started to publish technical standards which could help mitigate algorithmic bias in the design and deployment of AI systems. In April 2016 the British Standards Institute published BS 8611 *Guide to the Ethical Design and Application of Robots and Robotic Systems*, which is almost certainly the first explicit ethical standard in robotics and automation.[83] It provides an overview of 20 distinct ethical hazards and risks, grouped under four categories: societal, application, commercial and financial, and environmental.[84] Advice is given on mitigating the impact of each risk, along with suggestions of how these measures might be verified or validated. A range of organisations including the Institute for Electrical and Electronics Engineers (IEEE and its standards association IEEE-SA) and the International Organization for Standardization (ISO), are also aiming to develop standards which cover algorithmic bias.

At the European level, the EU's standards bodies, the European Committee for Standards (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) created an official Focus Group on AI in December 2018, in order to support ISO/IEC SC42, a new Joint Technical Committee for the development of standards related to AI. The primary aim of this focus group is to deliver a roadmap for AI standardization by early 2020, with recommendations for international standardisation, EU technical committees and EU policymakers.

However, these initiatives are somewhat unusual in the sense that they have been driven by the EU and an international standards organisation, respectively. The majority of standardisation initiatives currently in development appear to be industry-led, primarily by the sectors in which these systems are being used. This includes internet companies, professional services, manufacturing, banking and transport. While academics have been involved in some of these

[81] Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. Science, Technology, & Human Values, 41(1), 93-117.
[82] Hildebrandt, M., & Koops, B. J. (2010). The challenges of ambient law and legal protection in the profiling era. The Modern Law Review, 73(3), 428-460.
[83] BSI-2016. (2016). BS 8611: 2016 Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems.
[84] The societal hazards include, for example, loss of trust, deception, infringements of privacy and confidentiality, addiction, and loss of employment. Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics*, *2*(2), 46.

initiatives, the extent to which standardisation efforts are engaging with civil-society groups is unclear, although the IEEE-SA standards initiative are making significant efforts in this regard.

At the other end of the deployment process, initiatives are also underway to develop certification programmes which will clarify when systems have been tested for algorithmic bias and the measures taken to mitigate identified biases. One notable example is the Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). This was launched at the end of 2018 by IEEE-SA, in collaboration with founding partners including the Finnish Ministry of Finance, the cities of Espoo, Vienna and New York, UN bodies such as Unicef, and the High-level Panel on Digital Cooperation, and industry partners such as EY and Accenture. The aim of ECPAIS is to create specifications for certification and marking processes that improve transparency and accountability, and help to reduce algorithmic bias in AI systems, with a first draft certification programme planned for the end of 2019.[85]

The number of initiatives aiming to develop technical standards and certification schemes which deal with algorithmic bias raises questions of its own. It is as yet unclear whether these schemes will align with one another, and should significant differences emerge, whether companies and organisations will need to make their own choices about which standards and certification programmes to adhere to, or whether some will gain more traction than others. The European Commission has acknowledged this point in its 2019 Rolling Plan for ICT Standardisation, which has a focus on ethics and safety.[86] It calls for the fostering of coordination on standardisation efforts in relation to AI across Europe, and for coordination with wider international standardisation efforts. It remains to be seen whether this occurs, and, on a broader level, given the early stages which most of these efforts are currently at, the extent to which they will prove effective in mitigating algorithmic bias.

[85] IEEE Standards Association. The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS): https://standards.ieee.org/industry-connections/ecpais.html [accessed on: 14/06/19].
[86] European Commission (2019). 2019 Rolling Plan for ICT Standardisation: https://ec.europa.eu/docsroom/documents/34521 [accessed on: 14/06/19].

4. Financial services

Chapter summary

- Algorithmic decision-making is already widespread in the financial services sector, particularly in the areas of credit scoring, mortgage and loan decisions, and increasingly in customer-facing roles as well.
- As such, there is significant potential for algorithmic bias to have a detrimental impact on the financial wellbeing of individuals—with respect to credit scoring, there is evidence that this is already causing discriminatory outcomes for particular groups of consumers.
- Identifying algorithmic bias in financial services can be especially challenging, due to the proprietary nature of the systems and datasets being used, which companies may be unwilling to open up to researchers, and the difficulties in accepting which kinds of consumer segmentation should be considered acceptable.

Financial services are one of the areas where algorithms are being used most intensively and where there has been most analysis of their impact.[87] Unsurprisingly, the industry has also been a focus of debate for those concerned about algorithmic bias. For example, according to Trevor Philips, the former chair of the Equality and Human Rights Council:

> "it is now a commonplace in artificial intelligence and machine learning that algorithms that govern mortgage lending [and] insurance quotes...are biased against women, and even more so against people of colour. Black people pay billions in extra premiums and higher loan rates... But no one quite knows how and why the machines learn to discriminate, much less how to stop them."[88]

Others, like the House of Commons Science and Technology Committee,[89] The Economist[90] and The Atlantic,[91] have raised similar concerns.

---

[87] O'Neill, C. (2016). *Weapons of Math Destruction*. UK: Penguin Books. For a more in depth review of the literature in relation to finance, see Buchanan, B. G. (2019). Artificial Intelligence in Finance. *Alan Turing Institute*: https://zenodo.org/record/2626454#.XQz8ttNKjBI [accessed on: 21/06/19].
[88] Philips, T. (2019). Dear Cambridge, if you truly want to atone for slavery…: https://www.thetimes.co.uk/article/dear-cambridge-if-you-truly-want-to-atone-for-slavery-qldpjthrt [accessed on: 14/06/19].
[89] Authority of the House of Commons (2017). Algorithms in decision-making: https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf [accessed on: 14/06/19].
[90] The Economist (2018). How an algorithm may decide your career: https://www.economist.com/business/2018/06/21/how-an-algorithm-may-decide-your-career [accessed on: 14/06/19].
[91] Kirchner, L. (2015). When Discrimination Is Baked Into Algorithms: https://www.theatlantic.com/business/archive/2015/09/discrimination-algorithms-disparate-impact/403969/ [accessed on: 14/06/19].

4.1 Background

These commentators may be right to sound the alarm about algorithmic bias in financial services. However, like the other sectors discussed in this chapter, there is a long history of bias and inequity in the provision of financial services. Furthermore, while the fears surrounding algorithmic bias may be relatively new, the use of algorithms in this sector (as opposed to more recent advances in the application of machine learning) dates back to the 1960s, and has become widespread and multifaceted in the decades since.[92]

Much of this evidence relates to inequalities experienced by ethnic minorities and women in accessing credit, either as business owners or individuals.[93] For example, a 2016 working paper published by the National Bureau of Economic Research found that, even after controlling for credit scores and other risk factors, African-American and Hispanic borrowers in the United States were considerably more likely to have high-cost mortgages than Caucasian borrowers.[94] Interestingly, however, the authors do not attribute the difference to algorithmic bias. Instead, they suggest that members of these groups are less likely to shop around for mortgage products and therefore more likely to pay higher prices.

This example helps underline an important challenge for those working in this field, the task of disentangling bias caused by an algorithm versus bias or unequal outcomes caused by other factors. As we will see, there is some direct evidence of algorithmic bias, but even in the case of financial services the literature on this subject is best described as emerging and incomplete.

4.2 Use of algorithms: examples and intensity

What is clear is that the utilisation of algorithms, and especially machine learning algorithms, in financial services has grown rapidly in recent years. One of the most widely known uses is algorithmic trading or High Frequency Trading, which grew from less than 10% of equity trades in the early 2000s to as high as 40% in 2016.[95]

Another area which has seen the often controversial use of algorithms is payday loan companies, many of whom have become notorious over the last decade for greatly expanding the sources of data used to make fast and direct loan decisions to customers excluded from

---

[92] Pardo-Guerra, J. P. (2010). Creating flows of interpersonal bits: the automation of the London Stock Exchange, c. 1955–90. Economy and Society, 39(1), 84-109.

[93] Carter, S., Mwaura, S., Ram, M., Trehan, K., & Jones, T. (2015). Barriers to ethnic minority and women's enterprise: Existing evidence, policy tensions and unsettled questions. *International Small Business Journal*, *33*(1), 49-69; Haughwout, A., Mayer, C., Tracy, J., Jaffee, D. M., & Piskorski, T. (2009). Subprime mortgage pricing: the impact of race, ethnicity, and gender on the cost of borrowing. *Brookings-Wharton Papers on Urban Affairs*, 33-63; Asiedu, E., Freeman, J. A., & Nti-Addae, A. (2012). Access to credit by small businesses: How relevant are race, ethnicity, and gender?. *American Economic Review*, *102*(3), 532-37.

[94] Bayer, P., Ferreira, F., & Ross, S. L. (2017). What drives racial and ethnic differences in high-cost mortgages? The role of high-risk lenders. *The Review of Financial Studies*, *31*(1), 175-205.

[95] Aldridge, I., & Krawciw, S. (2017). *Real-time risk: What investors should know about FinTech, high-frequency trading, and flash crashes*. John Wiley & Sons.

more traditional forms of credit. Rather than rely on conventional credit scores alone, these companies have often used algorithms to sift thousands of data points about applicants, ranging from the brand of computer they use to the connections they have on social media.[96]

Algorithmic systems are also being used across an increasingly wide variety of areas, including personalised finance, trading, fraud analysis, and robo-advice.[97] Typical uses of algorithmic systems in the financial services industry include:

- **Customer-focused uses:** credit scoring, insurance, client-facing chatbots, sentiment analysis, robo-advice
- **Operations-focused uses:** capital optimisation, model risk management & stress testing; market impact analysis
- **Trading and portfolio management:** trade execution
- **RegTech:** facilitating regulatory compliance and supervision, fraud detection

Specific examples of algorithmic systems used in financial services include:

- **Ulster Bank's CRM platform:** the "SalesForce Einstein" AI component in Ulster Bank's customer relationship management (CRM) platform combines machine learning, deep learning, predictive analytics, natural language processing and data mining techniques to allow the Bank to continually model its 1.9 million customers to improve business decision-making.[98]
- **ING's Katana**: ING bank launched a tool called Katana that uses "predictive analytics to help traders decide what price to quote when a client wants to buy or sell a bond."[99]
- **CLEO Chatbot:** the chatbot app CLEO uses artificial intelligence to give people advice on how to optimise their finances.[100]

---

[96] Deville, J. (2013). Leaky data: How Wonga makes lending decisions. *Charisma: Consumer Market Studies*.
[97] Carey, S. (2019). How UK banks are looking to use AI and machine learning: https://www.computerworlduk.com/galleries/data/how-uk-banks-are-looking-embrace-ai-machine-learning-3657529/ [accessed on: 14/06/19].
[98] Atos. Putting artificial intelligence at the heart of business: https://atos.net/en/customer-stories/ulster-bank [accessed on: 14/06/19].
[99] Williams-Grut, O. (2017). Banks are looking to use artificial intelligence in almost every part of their business: Here's how it can boost profits: https://www.businessinsider.com/ai-in-financial-services-2017-11 [accessed on: 14/06/19].
[100] Williams-Grut, O. (2017). Banks are looking to use artificial intelligence in almost every part of their business: Here's how it can boost profits: https://www.businessinsider.com/ai-in-financial-services-2017-11 [accessed on: 14/06/19].

- **Olivia Chatbot:** HSBC launched a chatbot, Olivia, for verifying customer identity in order to increase security via an individual "voiceprint".[101]

4.3 Evidence of algorithmic bias in financial services

Banks and other financial services companies use information about consumers to formulate credit scores on individuals.[102] These scores are used to judge how likely an individual is to pay back money lent to them, and thereby determine how likely an individual is to receive access to certain products or services, such as a mortgage. As credit is sought by a large portion of the population, this means that any algorithmic biases present in credit scoring systems could have a huge impact on society, and a disparate impact on different social groups.[103]

Lenders are required by law[104] to assess the 'creditworthiness' of customers[105] for affordability (i.e. whether the borrower will be able to pay back the debt) and to protect vulnerable borrowers.[106] Accurate and reliable scoring methods are essential tools for lenders, and the use of algorithmic decision-making tools based on large quantities of data is becoming increasingly popular.[107] However, there is a lack of transparency within this, for example as people are often unaware of how their credit scores are devised, and thus unable to contest unfair or biased results.

Algorithms used to decide credit scores have tended to be trained exclusively on successful applicants. However, this sample bias in algorithmic training data has been found to have a negative impact on the algorithm's credit scoring performance.[108] Research has found that models which were trained exclusively on 'successful' cases were comparatively less accurate in detecting 'rejected' cases. This indicates a need for algorithms to be trained on accepted and rejected cases in a stratified manner according to the relevant target population, to mitigate biases and inaccuracies in credit scoring algorithms.

---

[101] HSBC UK. Telephone Banking: https://www.hsbc.co.uk/1/2/voice-id [accessed on: 14/06/19].
[102] Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2-3), 59-88.
[103] Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale JL & Tech.*, *18*, 148.
[104] In the UK credit reference agencies are overseen and licensed by the Financial Conduct Authority, while in the US this is regulated under the Fair Credit Reporting Act and the Equal Credit Opportunity Act.
[105] Financial Conduct Authority. FCA Handbook: https://www.handbook.fca.org.uk/handbook/CONC/5/?view=chapter [accessed on: 14/06/19].
[106] Aggarwal, N. (2018). Law and Autonomous Systems Series: Algorithmic Credit Scoring and the Regulation of Consumer Credit Markets: https://www.law.ox.ac.uk/business-law-blog/blog/2018/11/law-and-autonomous-systems-series-algorithmic-credit-scoring-and [accessed on: 14/06/19].
[107] Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision-making. *Big data*, *1*(1), 51-59.
[108] Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, *54*(8), 822-832; Verstraeten, G., & Van den Poel, D. (2005). The impact of sample bias on consumer credit scoring performance and profitability. *Journal of the operational research society*, *56*(8), 981-992.

This is in addition to issues caused by inconsistencies in how different scoring methods are employed by different credit scoring companies.[63] As a result of these, there are likely to be different potential biases, as well as multiple forms of the same bias-type, across different service providers. There is currently a lack of systematic empirical research into the impact of biases in credit scoring system on different social groups. Nevertheless, with consideration towards biases which can be inherent in the data models are trained on, it may be inferred how easily this could perpetuate existing patterns of exclusion and discrimination.[109]

A further phenomenon highlighted by academics is that of 'creditworthiness by association'. In contrast to traditional forms of credit scoring which have focused almost exclusively on an individual's credit history, this can see forms of social data being used—where an individual has shopped, where they live or who they are friends with, for example. In one particularly discussed case from 2008, a man found that the credit limit on his card had been reduced from $10,800 to $3,800 because American Express has determined that 'other customers who ha[d] used their card at establishments where [he] recently shopped have a poor repayment history with American Express'.[110]

Ultimately, any process which attempts to distinguish between the value of particular customers will always raise difficult questions around fairness. One solution, as with Amazon's Prime same-day delivery service (see Box XX), is for companies to provide all their services to all consumers in an equal way, although this would seem to preclude so-called 'price discrimination' strategies, which long pre-date the use of algoirthms and in some contexts may be relatively uncontentious (for example, products and services which are offered at lower prices to students and those on low incomes).

Equally, entirely legal pricing strategies, which do not distinguish between customers on the basis of protected characteristics but instead focus on other characteristics, as when companies have targeted low-income households or vulnerable individuals for loans with excessively high interest rates, can still be highly ethically dubious. Determining whether a system or strategy is discriminating on the basis of sound economic analysis and some consideration of fairness, or on the basis of prejudiced assumptions, is sometimes possible but not always easy, and requires careful analysis of whether, after all other factors are excluded, a particular group is paying more for a product or service.[111]

4.4 Case study: Algorithmic Redlining

The tendency for algorithms used in financial services to compare consumers who are deemed similar in some way when determining the accessibility of cost of products has clear implications for the fair treatment of customers. A related phenomenon which has been

[109] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.
[110] Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale JL & Tech.*, *18*, 148.
[111] Dobbie, W., Liberman, A., Paravisini, D., & Pathania, V. (2018). *Measuring bias in consumer lending* (No. w24953). National Bureau of Economic Research.

identified by scholars is 'algorithmic redlining', a practice which takes its name from the earlier twentieth-century practice in which mortgage lenders in the United States would literally draw lines on maps to indicate where it was considered 'too risky' to lend to African Americans who were looking to buy property.[112] Such concerns have been present for decades, having been described as 'weblining' in the media as early as 2000.[113]

In its modern incarnation, the term generally refers to automated or semi-automated lending decisions (or, in broader definitions, decisions relating to the provision of any product or service) which may exclude or disadvantage particular groups based on their ethnicity. Various academics have argued that this is a significant risk if historical data on housing and lending decisions is used, as algorithms learn to copy previous patterns of discriminatory decision-making, on top of the previously mentioned algorithmic decisions around credit scoring more generally.[114]

There has also been discussion of 'reverse redlining', whereby particular groups—such as African Americans in the run-up to the 2008 recession—are specifically targeted by algorithms looking for individuals to sell high risk but profitable loans to.[115]

Tackling this issue is not a simple case of banning considerations of ethnicity in these decisions—while the United States passed the Fair Housing Act in order to address this in its non-algorithmic form, and the UK Equality Act would similarly render decisions of this nature being based on ethnicity, machine learning systems may pick up on proxies for ethnicity (including almost anything marginally more statistically associated with one ethnic group over another), and use these as the basis of decisions instead.[116] This can even occur in cases where the overall policy objectives appear benevolent, as in the case of semi-randomised algorithmic lottery systems used in New York City to allocate scarce affordable housing. In some cases these systems appeared to be prioritising families who were already living in an area, which critics claimed was a factor in keeping particular areas ethnically homogenous.[117]

---

[112] As noted by James Allen, even this earlier, pre-digital decision-making process can be understood as a very simple (non-digital) algorithm—'"area" plus "colored people" equals "do not lend."' Allen, J. A. (2019). The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining. *Fordham Urb. LJ*, *46*, 219.

[113] See Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review*, *16*(1), 18–84. https://doi.org/10/gdxthj, p. 29.

[114] Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 2053951717743530; Domino (2018). On Ingesting Kate Crawford's "The Trouble with Bias": https://blog.dominodatalab.com/ingesting-kate-crawfords-trouble-with-bias/ [accessed on 14/06/19].

[115] Fisher, L. E. (2009). Target marketing of subprime loans: Racialized consumer fraud & reverse redlining. JL & Pol'y, 18, 121.

[116] Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: challenges, solutions, and policy implications. Journal of Information Policy, 8, 78-115.

[117] Allen, J. A. (2019). The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining. *Fordham Urb. LJ*, *46*, 219.

As with many other examples of potential algorithmic bias, there are relatively few concrete examples of where algorithmic redlining may be occurring and why, often due to the closely-guarded details of the proprietary algorithms being used and the individualised nature of the decisions being made. However, one striking example noted by academics in recent years was Amazon's so-called 'prime-lining scandal', where it was found that when Amazon made free same-day delivery available to Prime service subscribers in the US, only some areas were eligible—customers in predominantly African American residential areas were excluded, sometimes even when they closely bordered predominantly white areas which were included.[118]

Amazon denied accusations of racial profiling, and revealed that they had taken a largely automated, data-driven approach, which had only looked to prioritise the service in areas where there were already high densities of Prime subscribers, as this would prove most cost-effective for them; this had, however, often closely correlated with racial divides in US cities for long-engrained socio-economic reasons. The fact that, following a backlash, Amazon subsequently chose to disregard its algorithm, and make free same-day delivery available across all areas of these particular cities, illustrates that sometimes, fairness may mean a company choosing to make a less profitable decision.

4.5 Challenges and gaps

- The extent and scale of the problem in financial services is still relatively unclear—decisions around finance and credit are often highly opaque for reasons of commercial sensitivity and competitiveness. Even where discrepancies in outcomes are apparent, without extensive access to the models used by companies in their assessments of individuals, attributing bias to any specific algorithmic function is very difficult.
- Increasingly data-savvy investigative journalism has produced some startling examples of algorithmic bias in financial services, but there are still very few systematic studies comparing the treatment of individuals from different backgrounds by algorithmic systems with regard to financial decisions. This is especially the case in the UK and Europe.
- Extensive evidence also suggests that algorithmic approaches may help counter pre-existing human biases in the provision of financial services.[119] However, this may prove difficult if the suppliers of financial products do not feel inclined to make their data and algorithmic models available, or if insufficient consumer data is available to make accurate assessments of their needs. More generally, smaller numbers of training datasets for minority communities might result in the reduced performance of investment advice algorithms for these communities.

[118] Bruckner, M. A. (2018). The promise and perils of algorithmic lenders' use of big data. *Chi.-Kent L. Rev.*, *93*, 3.
[119] Baker, T., & Dellaert, B. (2017). Regulating robo advice across the financial services industry. *Iowa L. Rev.*, *103*, 713.

5. Local Government

Chapter summary

- There is evidence that algorithmic decision-making systems are being deployed by local governments for a wide variety of purposes, with particular interest in systems for the prediction of resourcing needs and identifying future risks.
- There is very little oversight of how councils are using these technologies, and the degree of public transparency also varies considerably across councils, making it difficult to map the diffusion of these approaches nationally, or determine how appropriately they are being used.
- Biases may be inherited from the commercial datasets and systems which councils are using, and for the time being the discussion over mitigation of bias appears to be less well developed in the UK as compared to some parts of the US.

The possible uses of algorithms for assisting decision-making in local government are many and varied, and include improving administrative office efficiency, prioritising assistance for homeless people, targeting housing inspections, alerting relevant authorities to possible child welfare issues (see case study below) and helping local residents with enquiries about local services.[120]

5.1 Background

Given the resource demands that many local councils face, a number of commentators have predicted that the use of algorithms will only continue to grow. The Transformation Network, which ranks local councils in England by their willingness to embrace new technology, argue:

> "Local government is facing a perfect storm … The best way to protect the future of local council services and the communities is through the smart use of technology, such as robotic process automation and AI. Far from something to be feared, such technology can liberate employees from mundane, repetitive work and allow them to spend more time doing what people do best, and that's providing front-line services to citizens."[121]

[120] Toros, H., & Flaming, D. (2018). Prioritizing Homeless Assistance Using Predictive Algorithms: An Evidence-Based Approach. *Cityscape*, *20*(1), 117-146.; Bright, J., Ganesh, B., Seidelin, C., & Vogl, T. M. (2019). Data Science for Local Government. *Available at SSRN 3370217*.
[121] Transformation Network. Local Government League Table: https://transformationnetwork.co.uk/local-government-league-table/ [accessed on: 14/06/19].

However, as local government begins to deploy algorithms in areas traditionally reliant on human judgement and sensitivity to make decisions affecting the welfare of citizens, there are also risks that algorithmic bias could lead to greater unfairness in the provision of public services.

Local government use of algorithmic decision-making systems is still relatively emergent, and as such, there are few detailed empirical assessments of working systems. Instead, the few studies that have been undertaken have generally relied on interviews with practitioners to ascertain some of the potential and emerging issues in this area.[122]

5.2 Uses of algorithms: examples and intensity

Assessment of the impact of algorithmic decision-making systems on local government in the UK varies considerably. While one recent survey based on local government responses found that less than 5% of English local councils said they were currently undertaking AI-related projects, another project which used Freedom of Information requests to gather information found that at least 53 councils were using algorithms in a predictive capacity (see below for more detail).[123] Where the real answer lies is probably as much a case of what precisely is being defined and counted; nevertheless, the evidence clearly suggests that algorithmic decision-making is in use today and is likely to expand in the future.

One of the most significant areas of interest for councils is in making predictions about the future needs and risks associated with local residents. These systems might, for example, aim to predict risk factors, where citizens and services might be at risk (having a relative in prison, for example, is a risk factor for a child in care becoming homeless) and protective factors, where positive outcomes can be understood and encouraged (having a good education, for example, can help protect an individual from negative outcomes).[124]

Probably the most extensive investigation into the use of predictive algorithms in local government in the UK has been conducted by Cardiff University's Data Justice Lab.[125] They found that local councils were using predictive algorithms for a wide variety of purposes. Camden Council was attempting to merge data to provide a 'single view of the citizen', and detect fraud; Kent was using data analytics for public health analysis; Bristol's Integrated Analytical Hub was attempting to predict the risk of child exploitation; and Hackney Council was

---

[122] Veale, M., Van Kleek, M., & Binns, R. (2018, April). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 440). ACM.
[123] Eichler, W. (2019). 'Shockingly small' number of councils embrace automation, study reveals: https://www.localgov.co.uk/Shockingly-small-number-of-councils-embrace-automation-study-reveals/47387 [accessed on: 14/06/19]; Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report.
[124] Selwyn, R. (2018). Predictive Analytics: https://troubledfamilies.blog.gov.uk/2018/05/14/predictive-analytics/ [accessed on: 14/06/19].
[125] Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report. *Data Justice Lab.*

partnering with companies Xantura and EY to use predictive analytics to identify children and families in need of intervention and extra support.

In Hackney, the local council developed its 'Early Help Predictive System' to analyse data related to (among other factors) debt, domestic violence, anti-social behaviour, and school attendance, to build a profile of need for families. By taking this approach, the council believes they can intervene earlier and prevent the need for high-cost support services later and further down the line. The new system is identifying 10-20 families a month who may need the attention of local services for these reasons.

A related area of algorithmic decision-making by local governments is in the planning sphere. For example, the 'Data Science for Social Good'[126] initiative of the University of Chicago has been using machine learning to help the City of Rotterdam[127] understand their rooftop usage to address challenges with water storage, green spaces and energy generation. Data Science for Social Good also helped the City of Memphis[128] map properties in need of repair.

There are also a number of examples of councils using AI-style technology in customer facing operations. For example, Milton Keynes Council has developed a virtual assistant (or chatbot) to help respond to planning-related queries. As part of its Connected Knowledge programme, Aylesbury Vale District Council has implemented a programme to integrate AI into its customer service operation.[129]

5.3 Evidence of algorithmic bias in local government

Much of the analysis that exists to date in the UK local government context has been concerned with promoting the use of data-driven approaches,[130] and as a result has not focused attention on potentially more negative considerations such as algorithmic bias. However, a more critical literature is beginning to emerge within the field of journalism. Following in the footsteps of US investigative journalism, which has identified data-orientated local government reform as a key area of emerging interest,[131] academics and journalists in the UK are starting to take a more critical look at the use of data by local government, and are highlighting the need for

[126] Data Science for Social Good: https://dssg.uchicago.edu/projects/ [accessed on: 14/06/19].
[127] Data Science for Social Good. Identifying rooftop usage in Rotterdam: https://dssg.uchicago.edu/project/identifying-rooftop-usage-in-rotterdam/ [accessed on: 14/06/19].
[128] Caro, A., Conway, M., Green, B., Manduca, R., & Plagge, T. Easing the distress of neighbourhoods with data: https://dssg.uchicago.edu/2014/11/12/easing-the-distress-of-neighborhoods-with-data/ [accessed on: 14/06/19].
[129] Digital Genius. Customer Service Automation Platform: https://www.digitalgenius.com/ [accessed on: 14/06/19].
[130] See for example Nesta's 'Local Datavores' project. Symons, T. (2016). Wise Council: Insights from the Cutting Edge of Data-Driven Local Government.
[131] See, for example, the Algorithmic Tips project. Trielli, D., Stark, J., & Diakopoulos, N. (2017). Algorithm Tips: A Resource for Algorithmic Accountability in Government. *Computation & Journalism Symposium*.

transparency in algorithmic processes to better ensure fairness and accountability.[132]

As mentioned, the Data Justice Lab, based at Cardiff University's School of Journalism, Media and Culture, has conducted one of the only studies attempting to systematically document the use of algorithmic prediction systems by local councils in the UK. Across these systems they found significant variation in the amount of information local councils are prepared to reveal about the systems they use. In some cases, they found councils either used vague terminology, or withheld information completely, in response to Freedom of Information requests. There were also no standard practices or common approaches with regards to how data is or should be shared and used by local authorities and partner agencies. While some councils were developing systems in-house, many others were using privately-developed systems; of these, Experian's Mosaic geo-demographic tool featured particularly prominently. Xantura, Callcredit and Capita also routinely provide data sharing and analytics services.[133] Large tech companies are also starting to develop business partnerships with local government—for example, Google's Better Cities Better Cities programme has been partnering with cities such as Amsterdam to show how anonymised, aggregate data from its Android mobile phone platform can be used to understand mobility patterns on its road network.[134] Any issues with bias in these products and datasets are therefore likely to carry over into the local council's uses of these systems.

There is evidence that at least some of the councils deploying algorithmic decision-making systems are actively considering the question of bias and how it might be mitigated in practice. To return to the example of Hackney Council, the Data Justice Lab quoted a document from London Councils and EY, which noted that "over 80% of households in Hackney that have been identified most at risk by the model are at risk".[135] They also cited the developers of the system who highlighted its ability to monitor for bias, with one developer explaining:

> "[I]f we look at the child protection caseload, we can see, for example, what the age distribution of the children are, I can see what the ethnicity distribution for this client is, I can see what all those different characteristics are, what the deprivation is. For each of those distributions, I can say what does the model do, is it a different distribution to the distribution in the client's caseload? If it's significantly different and the model is skewing oddly, why has it got a different bias to what's naturally in the data? This could be because there is bias the client's existing system or that the model is biased? So for the first time, I think, we can actually start looking at bias in the system, which is actually quite a powerful thing to be enabling."

---

[132] Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 440). ACM.
[133] Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report. *Data Justice Lab.*
[134] Bright, J., Ganesh, B., Seidelin, C., & Vogl, T. M. (2019). Data Science for Local Government. *Available at SSRN 3370217*.
[135] Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report. *Data Justice Lab.*

However, for external observers to identify potential bias, in addition to the system's success rate, we would also need to know the rate of false positives, e.g. how many people were incorrectly identified as high risk. The Data Justice Lab also argue that these kinds of comparisons will not address the extent to which other biases may enter a system, such as those caused by assumptions about what is 'normal' and what a functional family should look and act like.[136]

As with other areas of deployment, there is also a risk that algorithmic biases may take the place of implicit human biases, and it has been noted that there is the possibility of specific communities or social groups being discriminated against. For example, individuals who receive welfare payments are likely to have more data kept on them simply as a result of receiving this assistance, and thereby may be flagged as being higher risk and be investigated more often, creating a compounding bias loop as a consequence. Furthermore, the recommendations put forward by the algorithm are then considered by human caseworkers, who may not be sufficiently trained to understand and evaluate the impact of algorithmic bias.

Despite these issues, for the time being at least, local governments in the UK are developing these technologies in a mostly ad hoc fashion, and with comparatively little in the way of coordination or oversight. However, if developments in the United States are any indication of future trends in the UK, this may soon begin to change.

In New York City, for example, these concerns culminated in the passing of Local Law 49 in 2017, which mandated that city agencies produce lists of publicly-used algorithms (defined as 'automated decision systems' (ADS), which would be archived along with the data they use, and be auditable for issues such as bias. It also mandated the creation of an Algorithms Task Force, an independent body tasked with inspecting them.[137] While lauded and copied as a model for local algorithmic accountability across the United States and internationally, recent reports illustrate how difficult this process has been. After two years, city officials have struggled in practice to apply the broad definitions included in the law to operational systems, and no single instance of an ADS has been produced for inspection by the Task Force, despite the creation of an extensive list of possible examples by the AI Now Institute, an AI-focused NGO.[138] Members of the Task Force have accordingly begun to publicly question whether the process will produce any meaningful outcomes.

---

[136] Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report. *Data Justice Lab.*

[137] Stoyanovich, J., & Howe, B. (2018). Follow the data! Algorithmic transparency starts with data transparency.

[138] Lecher, C. (2019). New York City's algorithm task force is fracturing: https://www.theverge.com/2019/4/15/18309437/new-york-city-accountability-task-force-law-algorithm-transparency-automation [accessed on: 14/06/19]; Budds, D. (2019). New York City's AI task force stalls: https://ny.curbed.com/2019/4/16/18335495/new-york-city-automated-decision-system-task-force-ai [accessed on:14/06/19]. For AI Now's list of possible example of ADSs, which was prepared in advance of public forums hosted by the Task Force, see: AI Now, Automated Decision Systems: https://ainowinstitute.org/nycadschart.pdf [accessed on: 14/06/19].

5.4 Case Study: Child Welfare

In the US, algorithms have been used in order to assist in decision-making by human caseworkers with regard to child welfare. There are millions of referrals made each year to US child protective services, which are screened by the relevant jurisdiction to determine whether or not there will be a follow-up investigation.[139] The local nature of these decisions can however create large variations in how a similar referrals are treated in different parts of the country. Algorithmic implementation has been suggested in order to help standardise these decisions, as well as to provide predictive analytics to human screeners in order to help them assess referrals faster, more accurately, and with less potential for human biases.[140] Government austerity has also been cited as a factor behind the growing interest in algorithmic decision-making systems.[141]

In the UK, local councils are embarking on similar experiments in the area of child welfare. A 2018 newspaper investigation found that at least five local authorities have developed or implemented a predictive analytics system for child safeguarding. Hackney and Thurrock councils have both hired Xantura to develop a predictive model for their children's services teams. Two other councils, Newham and Bristol, have developed their own systems internally. Brent council is developing a system to predict vulnerability to gang exploitation. Collectively, these systems were found to draw on the data of at least 377,000 people.[142]

Bristol City Council's Integrated Analytical Hub has been developed in-house, a decision partly taken to ensure they fully understood the system they were using, which might not be the case if they relied on an external contractor's system with all the proprietary lack of transparency that could entail.[143] The Hub initially grew out of the council's work in the context of its Troubled Families programme, before they started to look at whether the same system could predict the risk of child sexual exploitation. Their predictive model now draws on data from 35 different social issue datasets, and also draws on some data from Experian. The model produces an automated risk score from 0 to 100 for every young person in the database, and this in turn informs the priorities of case workers and intervention teams. Work is now being done to expand the system beyond sexual exploitation, into a broader

---

[139] Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision-making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134-148).

[140] Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision-making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134-148).

[141] Bright, J., Ganesh, B., Seidelin, C., & Vogl, T. M. (2019). Data Science for Local Government. *Available at SSRN 3370217*.

[142] McIntyre, N., & Pegg, D. (2018). Councils use 377,000 people's data in efforts to predict child abuse. *The Guardian*. [Accessed on 25/06/2019] https://www.theguardian.com/society/2018/sep/16/councils-use-377000-peoples-data-in-efforts-to-predict-child-abuse

[143] Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report. *Data Justice Lab.*

'vulnerability index' which addresses other risks such as criminality and drug addiction.

Significantly, the system is framed as a decision-assisting tool. The algorithmic model only accounts for 'negative' data (factors which could increase the risk of exploitation, such as previous incidents of domestic abuse), and does not factor in 'insulating' factors which could decrease the risks (such as active engagement with social groups)—for this, the knowledge and experience of professional caseworkers is relied upon.

Bristol City Council do appear to have started considering issues of bias in relation to their system. The database it relies upon is updated is updated every week with an accompanying risk analysis, along with a corresponding accuracy measure. This warns users if the accuracy measure drops below a given threshold, in which case the database needs to be rebuilt.

However, there are several factors which could increase bias within the Bristol system. Some of the datasets the system uses have high error rates (such as arrest records, in which incorrect names or dates of birth are commonplace). While Bristol City Council's data scientists try to correct for this by, for example, prioritising data from more reliable datasets, actually correcting the errors is difficult, as the Analytical Hub team does not own or control the data which is shared with them by other parts of the council. The system is based on collecting as much data as possible and cleaned data cannot be passed back to the original data-set used as a source.

There are also questions of bias related to the way human users interact with the system. The team behind the system have tried to account for automation bias (where people begin to defer to automated decisions as standard, believing them to be more objective and less fallible than they are) by, for example, not using colours like red, amber and green, or to name something 'high risk' on the system. Data scientists behind the project have also highlighted risks of feedback biases emerging, if users do not understand what contributes to particular scores, with one observing to researchers:

> "You've got to be careful you don't end up generating some feedback loops where your scores feeds another score feeds your score, and you end up just constantly multiplying everybody's score each week. There's definitely a risk of that."[144]

In the US, similar concerns have even led to the cancellation of comparable programmes. In 2017, for example, both the Illinois Department of Children and Family Services and the County of Los Angeles Office of Child Protection terminated the use of predictive analytics programmes, in part due to perceptions of inaccuracy (including high false positive and false negative rates), the poor quality of data being used and the difficulty of verifying their decisions.[145]

[144] Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report. *Data Justice Lab.* 34.
[145] Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward Algorithmic Accountability in Public Services. CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK.

5.5 Challenges and gaps

- The current literature has mostly focused on the US context, and there are relatively few studies which focus on issues of algorithmic bias in the context of local government in the UK or at the European level, where governance structures and circumstances are often quite different.

- So far, no empirical studies of algorithmic systems and their observed biases within the context of local government appear to have been conducted. Where evidence does exist, it is usually reliant on journalistic or interview-based methodologies, with information largely coming from surveys, Freedom of Information requests and the analysis of technical documentation where available.

- Local governments often make use of a large variety of datasets, drawn from disparate sources and agencies. The sourcing and combination of datasets without a full understanding of the context, strengths and limitations of the data in question is understood to be a common factor which contributes to algorithmic bias, but a discussion of the ways in which data is sourced, shared, used and interpreted in practice by local governments and state agencies is lacking in current literature.

- Data science skills and expertise is beginning to develop within local government, but resource constraints mean that much of this is expertise is currently directed towards the meeting of statutory reporting requirements, and not the development of new systems. It is not clear what level of understanding local government officials more generally have regarding the issues of algorithmic bias.

6. Crime and Justice

Chapter summary

- The use of machine learning decision-making systems within policing, is still in its infancy; nevertheless, the use of predictive policing systems, which attempt to forecast crime trends across areas or generate individual risk profiles, is increasing. Facial recognition trials are also underway in some UK police forces.
- Predictive policing systems are heavily reliant on the use of historical crime data, and as such pose real risks of learning from historical biases prevalent within the criminal justice system.
- The facial recognition systems being trialled by forces also likely suffer from the kinds of biases detected in commercial systems, which can make them substantially less accurate in relation to ethnic minority groups. However, the lack of data being collected on ethnicity during these experiments makes this difficult to audit.

Algorithmic decision-making tools are increasingly being used in the areas of law enforcement and justice to help predict risks, prioritise resources and promote more consistent, evidence-based decision-making. The modern interest in predictive policing programmes in particular is usually traced back to the 2000s, and particularly the period after 2008, when many police forces in the US and the UK sought to maintain levels of service with limited resources.[146]

6.1 Background

Before going into further detail it is important to make a distinction between decision-*making* tools (such as predictive policing tools) and tools purely used to help officers act on human-made decisions (such as facial recognition or Automatic Numberplate Recognition (ANPR) technology), which might be thought of as 'decision-*assisting*' tools. Facial recognition and ANPR are used to flag possible matches with individuals that the police have already decided are of interest, through comparison to image and video evidence. Although new facial recognition systems are raising questions around privacy and algorithmic bias in the context of accuracy (discussed further below) the basic principles underpinning these technologies are relatively well established. These technologies are arguably different to predictive policing tools, which are able to infer connections between data (for example, drawing connections between

[146] Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation; Ferguson, A. G. (2016). Policing predictive policing. *Wash. UL Rev.*, *94*, 1109.

the times, locations, types of crimes in historic datasets to infer the probable trends in future crime) and subsequently, provide advice on what actions officers should take.

Predictive policing can be divided into two categories:

- Projects which aim to predict general future trends in crime, usually in terms of when and where particular crimes are likely to occur, in order to assist with resourcing and deployment decisions.
- Projects which aim to predict the risks of particular individuals committing offences, usually in the context of granting bail decisions, diverting certain offenders towards non-custodial sentencing options, or probation assessments.[147]

Tools which aim to highlight future crime hotspots and trends based on historical crime data are becoming increasingly commonplace, with the most well known example being a system developed by the US company PredPol. In the UK, Kent Police trialled this software for five years, before ending their contract in 2018, citing an interest in developing their own alternative system.[148] Perhaps the most commonly cited example of the latter category of systems, focused on predicting the risks associated with particular individuals (usually of offenders who may re-offend) is the COMPAS system, deployed in a number of places across the US (see Box XX).

Police forces have also been experimenting with a wide variety of algorithmic decision-making tools for making other aspects of their work more effective and efficient. Trials with facial recognition systems, which combine high-resolution cameras with an algorithmic decision-making system to scan and cross-reference faces against databases of images, either in real-time or after the event (such as scanning CCTV footage to identify suspects or witnesses) have been conducted in various parts of the world, and have attracted considerable academic interest.[149] But there has also been considerable growth in the use of algorithmic approaches to the processing of administrative police data, and the sifting of digital forms of evidence, which may also prove to have a significant impact on the shape of law enforcement in the future.

6.2 Uses of algorithms: examples and intensity

As with other sectors, judging the diffusion of algorithmic decision-making tools across law enforcement in the UK is partly a question of definitions, and even establishing where trials are occurring is not always easy in a highly decentralised system, with 45 separate territorial police

[147] National Institute of Justice (2014). Predictive Policing: https://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/Pages/welcome.aspx [accessed on: 14/06/19].
[148] Financial Times. First UK police force to try predictive policing ends contract: https://www.ft.com/content/b34b0b08-ef19-11e8-89c8-d36339d835c0 [accessed on: 14/06/19].
[149] Introna, L., & Nissenbaum, H. (2010). Facial recognition technology a survey of policy and implementation issues.

forces. However, it is clear that a significant minority of forces are experimenting with predictive policing programmes, while relatively few are embarking on facial recognition trials.

In evidence provided to a 2018 House of Commons select committee inquiry into algorithmic decision-making, Marion Oswald, Director of the Centre for Information Rights, and Sheena Urwin of Durham Constabulary, noted that only 14% of UK police forces were using algorithmic data analysis or decision-making for intelligence work. By contrast, research conducted by Liberty in 2019 showed that at least 14 police forces (around one third of forces in the UK) are currently using algorithmic predictive policing programs, have previously used them or are engaged in relevant research or trials.[150]

In some cases, deployment of predictive crime mapping systems UK forces can be traced back further than their US counterparts, with ProMap, developed by researchers at the Jill Dando Institute of Crime Science at UCL, first being deployed and evaluated by the Home Office and East Midlands Police in 2005 and 2006.[151] In the years since, West Midlands, Avon and Somerset, Kent, West Yorkshire, Norfolk and the Metropolitan Police have all developed, trialed or actively deployed predictive mapping systems.

Similarly, the use of algorithms in individual risk profiling and assessment is also well established in the UK, generally in the context of the prison and probation service, as well as, to a lesser extent, sentencing decisions. The Offender Assessment System (OASys) is a national risk/need assessment tool with algorithmic components used across probation areas and prison establishments in England and Wales. The system has been developed in various guises from the mid-1990s, and one of the most important components, OGRS4, which produces a score concerning the risk of reoffending, was trained on 1,809,000 offenders released from custody or disposed of otherwise between April 2005 and March 2008, who had not reoffended before the end of March 2008. In 2010 it was recalibrated based on a further set of 174,000 offenders.[152]

OASys provides three statistically validated indicators of reoffending, focusing on an individual's general risk of committing non-violent, non-sexual offences, their risk of committing violent offences, and their risk of reconviction.[153] Taken together, these indicators use a combination of static data (such as demographic data and criminal history) and dynamic data (such as data on lifestyle, 'thinking and behavior' and 'attitudes') to create risk scores that will inform the

[150] Couchman, H. (2019). Policing by Machine: https://www.libertyhumanrights.org.uk/sites/default/files/LIB%2011%20Predictive%20Policing%20Report%20WEB.pdf [accessed on: 14/06/19].
[151] Law Society (2019). Algorithms in the Criminal Justice System: https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/ [accessed on: 10/07/19]
[152] Law Society (2019). Algorithms in the Criminal Justice System: https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/ [accessed on: 10/07/19]
[153] Law Society (2019). Algorithms in the Criminal Justice System: https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/ [accessed on: 10/07/19]

prioritisation of offenders during and after their sentences, and, to a lesser extent by informing pre-sentencing reports, their actual sentence itself.

There are also some examples where crime mapping and individual risk profiling are integrated into a single platform, as with the case of the Qlik Sense system by Avon & Somerset Police. Originally envisaged as a simple data visualisation and management tool by the force in 2016, it has since evolved to cover 30 different applications across the force, and is now routinely used to develop risk profiles for around 250,000 individual offenders, which in turn is used to inform the allocation of operational resources.[154] It is also used for geographic crime mapping and predictions, making it one of the more wide-ranging examples of predictive policing in the UK.

Experiments with facial recognition systems have so far been much more piecemeal, with only three forces recently trialing the technology—Leicestershire Police, South Wales Police and the Metropolitan Police. They have deployed automatic facial recognition technology at shopping centres, festivals, sports events, concerts, community events and political demonstrations.[155] These deployments have attracted considerable public scrutiny, and the use of the technology in both South Wales and London are currently undergoing judicial review, which will have implications for the future use of facial recognition by UK forces.

6.3 Evidence of algorithmic bias

Notably, algorithmic bias was already being considered in the UK criminal justice system in the 2000s, before the recent growth in academic interest in the subject—when it was realised that the predictive validity of the OASys system was weaker for female offenders than for male offenders, the decision was taken to model age separately for each gender.[156] However, more recently the use of algorithmic decision-making systems in law enforcement has been heavily criticised, perhaps more so than any other sector considered in this review, and much of this criticism has focused on both the current reality and growing potential for these systems to arrive at biased decisions which may significantly impact lives.

Big Brother Watch have flagged concerns over potential biases in the facial recognition systems currently being trialled by some UK forces. In 2018 they identified high rates of misidentification (where an individual was inaccurately identified as a possible person of interest—i.e. a false positive), averaging around 95% across all trials,[157] although this is beginning to change, with

---

[154] Dencik, L., Hintz, A., Redden, J., & Warne, H. (2018). Data scores as Governance: Investigating uses of citizen scoring in public services project report.
[155] Big Brother Watch (2018). Face Off The lawless growth of facial recognition in UK policing: https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf [accessed on: 14/06/19].
[156] Law Society (2019). Algorithms in the Criminal Justice System: https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/ [accessed on: 10/07/19]
[157] Big Brother Watch (2018). Face Off The lawless growth of facial recognition in UK policing: https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf [accessed on: 14/06/19].

South Wales police reducing their rate of false positives to 50% in recent deployments.[158] For operational purposes, this is a problem in of itself, but Big Brother Watch also noted that given broader evidence that facial recognition systems are more accurate for White individuals compared to ethnic minority individuals (most likely stemming from larger training datasets for the former category),[159] there is a risk these mistakes are not affecting all people equally. Furthermore, the Metropolitan Police have not collected data on the ethnicity of individuals identified during their trials, on the grounds that this would be unnecessary and unimportant, therefore making it impossible to know whether their system was disproportionately affecting BAME individuals.[160]

A frequent criticism of the police use of predictive algorithms for mapping future crime trends is that the data being fed into these systems is biased from the very beginning, in ways which reflect how crimes are reported to, or detected by the police, as much as, or more than, the reality of crime in a given area. To take the example of an apparent spike in commercial burglaries being reported between 7 and 8 am—it may be difficult to determine from data alone whether this was when burglaries were actually occurring, or simply when property owners and managers discovered and reported burglaries which took place overnight.[161] Liberty's recent report on the subject argued similarly:

> "Using historical crime data to make predictions is deeply problematic because the data collated by the police does not present an accurate picture of crime committed in a particular area—it simply presents a picture of how police responded to crime".[162]

Lum and Isaac have noted that if the police focus their attention on specific areas or social groups, these may become systematically over-represented in the data.[163] In practical terms, if they police an area more frequently, they are correspondingly more likely to detect crimes there, thus creating a cycle of compounding bias as these areas appear more frequently in the database. Other researchers have raised similar concerns. One simulation-based study found

---

[158] Davies, B., Innes, M., & Dawson, A. (2018). An Evaluation of South Wales Police's use of Automated Facial Recognition: https://static1.squarespace.com/static/51b06364e4b02de2f57fd72e/t/5bfd4fbc21c67c2cdd692fa8/1543327693640/AFR+Report+%5BDigital%5D.pdf [accessed on: 14/06/19].

[159] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77-91); Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789-1801.

[160] Big Brother Watch (2018). Face Off The lawless growth of facial recognition in UK policing: https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf [accessed on: 14/06/19].

[161] Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation; Ferguson, A. G. (2016).

[162] Couchman, H. (2019). Policing by Machine: https://www.libertyhumanrights.org.uk/sites/default/files/LIB%2011%20Predictive%20Policing%20Report%20WEB.pdf [accessed on: 14/06/19].

[163] Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, *13*(5), 14-19.

that the use of predictive policing could trigger 'runaway feedback loops', whereby the inputting of detected crime data leads to additional police deployments in a particular area, who then detect further crimes and in turn feed even more data about crime in that area back into the system.[164]

Academic research and investigative journalism, especially in the US, has also highlighted the fact that predictive policing systems are often trained on compromised or flawed data, sometimes referred to as 'dirty data'.[165] Data fed into predictive policing algorithms may come from periods of history marked by flawed, racially biased and unlawful police practices that include systemic data manipulation, falsification of police reports, unlawful use of force and planted evidence. This has been shown to be an issue in some US cities with a recognised history of unlawful police practices, such as Baltimore, Chicago and New Orleans.[166] Andrew Selbst argues that, although predictive policing is portrayed as a neutral medium that counteracts unconscious biases, the presence of 'dirty data' means that systematic biases are "an artifact of the technology itself, and will likely occur even assuming good faith on the part of the police departments using it."[167] The presence of 'dirty data' in predictive policing systems raises the risk that these systems will be inaccurate, skewed, or systematically biased and thereby perpetuate negative trends in police practice.

This is not a phenomenon new to algorithmic decision-making in policing, and has previously been a criticised aspect of so-called 'zero tolerance' or 'broken window' approaches to policing, which can lead to the increasingly intensive policing of particular areas as police officers not only respond to crime, but also detect it. However, there are fears that algorithmic approaches could lead to the less reflective use of such strategies in the future.[168]

One recurring theme across developments in the UK is that not all trials and deployments of algorithms have been open to scrutiny by researchers. The Ministry of Justice has been unusually proactive in this respect, publishing all the model weightings for its OASys system publicly, although there appears to have been limited scholarly assessment of this. They have also published predictive validity on a variety of demographic subgroups, data on offenders broken down by offence type, and in-house and peer-reviewed work analysing and explaining disparities in predictive performance by age, gender and ethnicity. The situation with regard to

---

[164] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*.
[165] Rashida Richardson, Jason M. Schultz and Kate Crawford (2019). Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review*, vol. 94:192, 193-223.
[166] Rashida Richardson, Jason M. Schultz and Kate Crawford (2019). Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review*, vol. 94:192, 193-223.
[167] Andrew D. Selbst (2018). Disparate Impact in Big Data Policing. *Georgia Law Review*, vol. 52, 109-195.
[168] Haskins, C. (2019). Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed: https://www.vice.com/en_us/article/xwbag4/academics-confirm-major-predictive-policing-algorithm-is-fundamentally-flawed [accessed on: 14/06/19].

police forces is more mixed however, and a recent Law Society report on the use of algorithms in the criminal justice system made clear that while some forces have invited at least a degree of academic scrutiny, others have not, making it sometimes difficult to determine where there may be issues with bias.[169]

6.4 Case study: Algorithmic risk assessments

In various parts of the world, systems are now being trialed and deployed to predict the risk of recidivism, and advise operational decisions. The most well known example of this is software called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), developed by Northpointe (now Equivant), and used by a number of courts in the US to generate risk assessments for defendants, which in turn are used as one factor by judges when determining sentences.[170] The software generates scores suggesting the probability of an individual failing to appear for trial or committing further pretrial offences, the probability of general reoffending after trial, and the probability of violent reoffending.

COMPAS has been heavily criticised on the grounds that it has contributed to biased decisions by US courts, and delivers decisions which are no better than untrained humans chosen at random.[171] Investigative journalism non-profit ProPublica claimed to have detected significant bias against ethnic minority defendants who were assessed using the COMPAS tool. It not only mislabeled White defendants as low risk more often, but also wrongly labelled BAME defendants as high risk twice as often as White defendants.[172] While the 137 questions used to derive COMPAS scores do not include questions about race, the precise scoring mechanism is not known, and the results suggest that other sociodemographic, behavioural, or lifestyle factors gathered in the questionnaire act as proxies for race, resulting in biases. This was a factor in the case of Loomis v Wisconsin, in which Eric Loomis, who had been sentenced to six years in prison on the basis of advice from COMPAS, attempted to challenge the outcome based on the systems lack of transparency and potential biases. The Wisconsin Supreme Court however ruled against Loomis, on the basis that his sentence would have been the same regardless.[173]

---

[169] Law Society (2019). Algorithms in the Criminal Justice System: https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/ [accessed on: 10/07/19]

[170] Skeem, J., & Eno Louden, J. (2007). Assessment of evidence on the quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). Unpublished report prepared for the California Department of Corrections and Rehabilitation. Available at: https://webfiles. uci. edu/skeem/Downloads. html.

[171] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

[172] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [accessed on: 14/06/19].

[173] Yong, E. (2018). A popular algorithm is no better at predicting crimes than random people. *The Atlantic*.

It should be noted however that this criticism has not been universally accepted, either by the company which developed it or some academics, who argued that for any given score the system in fact displayed similar degrees of accuracy for White and Black offenders.[174] It appears that these diametrically opposed conclusions are the result of different understandings of fairness within the system. These differences come down to whether the desired outcome is to optimise for 'true positives'—which will identify as many people as possible who are at high risk of committing a crime, but will generate more false positives (people unjustly classified as likely reoffenders) as well—or to deliver as few false positives as possible, which will also increase the rate of false negatives (likely re-offenders who are not classified as such).[175]

In part because the reoffending rates for Black and White offenders do in fact differ in the United States, it is mathematically likely that the 'positive predictive values' (the percentage of all 'positive' results which are in fact true) for people in each group will be similar while the rates of false negatives are not. For example, you may have a low false positive rate (reducing the number of people unjustly labelled as high risk), but this may lead to a decrease in true positives (more high-risk offenders being labelled as low risk), while leaving the overall positive predictive value unaffected. Ultimately, the trade-off between optimising true positives, or reducing false negatives, is a moral decision, and cannot be reconciled within a machine alone, though research suggests that there are certain key factors which are highly determinant of whether a person is likely to think a particular algorithmic decision-making system is fair or not.[176]

Applications in the UK

Alongside the OASys previously discussed in the context of prison and probation services, similar tools are being deployed by police in the UK. Durham Constabulary developed its Harm Assessment Risk Tool (HART), which uses data across 34 different categories—covering a person's age, gender and offending history—to rate people as having a low, moderate or high risk of reoffending over a two-year period. This in turn helps determine the eligibility of individuals for the force's Checkpoint programme, an alternative to prosecution aimed at reducing reoffending.[177]

---

[174] Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, *80*, 38.

[175] Spielkamp, M. (2017). Inspecting Algorithms for Bias: https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/ [accessed on: 14/06/19].

[176] Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018, April). Human perceptions of fairness in algorithmic decision-making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 903-912). International World Wide Web Conferences Steering Committee.

[177] Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, *27*(2), 223-250.

Within this system a value-judgement was made to deliberately calibrate the system to minimise high risk false negatives (i.e. offenders who are predicted to be relatively safe, but then go on to commit a serious violent offence), at the expense of generating more false positives (offenders who are classed as high risk, but probably would not have gone on to commit a serious offence).[178] Within this context, the Durham Constabulary felt that the risk of placing a high risk offender within the Checkpoint programme would be greater than denying a low risk offender access to this programme. While many people would probably agree with this decision, it must be assumed that only a relatively small group of people were involved in making it, even though the outcomes may affect a large number of people across society. It is therefore vital to consider who is involved within these decision-making processes in order to ensure a fair, deliberative outcome. This is an example of how balancing between false positives and false negatives to get a 'fair' outcome (or at least an outcome designed to produce the least adverse effects) can impact massively on the lives of people who come into contact with these systems.

This system has also come in for criticism: the tool has at certain points in its development used Experian's Mosaic tool, which utilises information such as postcodes to make recommendations, and these could become proxy variables for protected characteristics, thus resulting in biased decision-making. Furthermore, an early iteration of the system explicitly included categories specifying racial heritage, meaning the system may have used ethnicity data when producing advice on custody decisions (a protected characteristic under the Equality Act 2010).[179]

These examples indicate the necessity of mitigating algorithmic biases, and not assuming that algorithmically-provided recommendations are neutral, despite attempts to employ them in order to avoid unconscious human biases. There is also the risk of what the legal academic Bernard Harcourt terms 'the ratchet effect': the more an individual has reason to interact with the criminal justice system, whether by being stopped by police, arrested or taken to trial, the more data is generated about that individual, and the greater the subsequent police interest in that individual will be, which could make it more difficult for the individual in question to reform their ways and break the cycle.[180]

---

[178] Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'experimental' proportionality. Information & Communications Technology Law, 27(2), 223-250.
[179] Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, *27*(2), 223-250.
[180] Rowe, M. (2018). EXPERT COMMENT: AI profiling: the social and moral hazards of 'predictive' policing: https://www.northumbria.ac.uk/about-us/news-events/news/2018/03/ai-profiling-the-social-and-moral-hazards-of-predictive-policing/ [accessed on: 14/06/19].

6.5 Challenges and gaps

- The extent to which algorithmic decision-making systems are actually basing decisions on real-world patterns of crime, as opposed to patterns of reported or detected crime data, is not well understood, and there is a lack of empirical studies in this area.
- The long-term impact of using predictive policing systems, and the potential biases which may build up during and after deployment, are currently not well understood on an empirical level—only simulations of possible outcomes have been studied.
- The extent to which police facial recognition systems are replicating the racial biases observed in commercial systems is very difficult to assess, as UK forces engaged in trials do not appear to be collecting the data needed to verify this.

7. Recruitment

Chapter summary

- Algorithmic systems for sifting job applications appear to be becoming a routine aspect of at least the initial stages of recruitment. Applicants are also increasingly looking to optimise their applications to take advantage of the biases within these systems.
- While some believe it could help to tackle deep seated unconscious biases in human sifters, others point to evidence that, without proper consideration and design, these systems are just as likely to proliferate and exaggerate pre-existing biases against women and ethnic minority candidates.
- There is also evidence that even the jobs people are made aware of in the first place are being skewed by algorithmic biases, with, for example, certain kinds of higher-paid jobs shown more frequently to men than women.
- However, very few systematic empirical studies into existing systems and products have been conducted in this area, making it difficult to judge the true extent and severity of recruitment-related algorithmic bias.

Recruitment was where one of the first public reckonings with algorithmic bias occurred in the UK, when St George's Hospital Medical School in London decided to develop a computer system to screen their applicants in the 1970s. In 1988, the Commission for Racial Equality found the hospital guilty of racial and sexual discrimination: the programme developed by the hospital, which they trained on data about previous decisions made by human sifters, had denied interviews to as many as 60 applicants because they were women or had non-European sounding names.[181]

Today, the use of data-driven algorithms to assist employers with recruitment is growing, but our understanding about the impact of algorithms has not necessarily grown with it. While data on the use of such approaches across the UK is not readily available, a recent US report found that 55% of US human resource managers are planning to use AI over the next five years[182] Online professional networking site LinkedIn, probably the world's largest professional social network and recruitment portal with around 250 million active users,[183] offers employers

[181] Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, *33*(4), 111-117.
[182] Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-airecruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [accessed on: 14/06/19].
[183] LinkedIn. About LinkedIn: https://news.linkedin.com/about-us#statistics [accessed on: 14/06/19].

algorithmic rankings of candidates based on their 'fit' for job postings on its site. And startups such as HireVue[184] offer advanced AI-based functionalities like speech and facial expression analysis in videos to reduce the reliance of recruitment on resumes.

7.1 Background

One of the central debates currently is focused on whether algorithmic screening could help counter the biases (conscious and unconscious) of human sifters, who are required to rapidly screen out many applications at the earliest stage without thoroughly reading applications. This is particularly of interest with regard to increasing diversity in the workforce.[185] A report by McKinsey highlighted the example of one professional services company which used an algorithmic approach to sift the 250,000 applications it receives every year, and observed a 15% increase in the rate of women who passed through its screening process, compared with the previous manual approach.[186] However, there are others who believe these systems are simply perpetuating and aggravating pre-existing biases, as discussed later in this section.

In turn, there is also evidence that many job applicants are increasingly aware of the use of algorithmic sifting, and are seeking to game these systems to their advantage. A recent online poll of 6,551 people conducted by the recruitment agency Hays found that:

- 27% indicated they have already adapted their CV and online profiles to take advantage of automated systems
- 54% planned to do so in the next 12 months
- 19% had no plans to adapt.[187]

7.2 Use of algorithms: examples and intensity

Mapping the scale of algorithmic decision-making as part of the recruitment process in organisations is challenging because such Human Resources practices are generally not part of their public facing communications. However, in December 2017 the UK-based AI-orientated consultancy CognitionX stated they were tracking 300 HR-related tools that use machine learning; more than 100 of them were focused on recruitment.[188] They outline a range of areas of activity including:

- candidate sourcing;
- compatibility matching (using psychometrics or other forms of selection);

[184] HireVue. Better Hiring with AI-Driven Predictions: https://www.hirevue.com/ [accessed on: 14/06/19].
[185] Houser, K. (2019). Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making. *Mitigating noise and bias in employment decision-making (February 28, 2019)*, *22*.
[186] de Romree, H., Fecheyr-Lippens, B., & Schaninger, B. (2016). People analytics reveals three things HR may be getting wrong. *McKinsey Quarterly*.
[187] Hays (2018). 81% of jobseekers adapt their CV for algorithm screening: https://www.hays.com.au/press-releases/HAYS_1990835 [accessed on: 14/06/19].
[188] Jeffery, R. (2017). Would you let AI recruit for you?: https://www.peoplemanagement.co.uk/long-reads/articles/recruiting-algorithms [accessed on: 14/06/19].

- predictive analytics around new hire performance;
- full recruitment platforms focused on improving candidate experience;
- video interviewing.

Their report suggests a number of more specific advanced AI features that are being developed by these companies. These include:

- algorithms to reduce biased language when writing job descriptions, or remove any information that may indirectly disclose protected characteristics from job applications;
- chatbots that can directly interact with applicants;
- methods to assess interviewee suitability from their facial expressions; and
- algorithms that harvest social media information to detect supporting evidence for submitted job applications.

Several other companies around the world are developing related AI-based recruitment services.[189]

7.3 Evidence of algorithmic bias

Algorithmic bias in recruitment largely stems from the kinds of historical data which are being used to train algorithmic decision-making systems on the kinds of candidate a company or organisation is looking for. As Hetan Shah of the Royal Statistical Society informed a House of Commons inquiry into algorithmic decision-making, the approach is often to ask: "Here are all my best people right now, and can you get me more of those?"[190] But while recruiters may assume their system will focus on particular professional skills or characteristics, a machine learning algorithm is just as likely to pick up on the fact that many candidates were male, or had traditional European names, and optimise selection to focus on that, potentially reinforcing this over several stages of training.

Various examples of algorithmic bias in recruitment systems are becoming apparent, although there are relatively few studies which have assessed the extent of this bias empirically.
For example, concerns around age- and gender-based discrimination have been raised in the context of targeted job advertisements, where the line between deliberate digital targeting and unintended algorithmic discrimination is unclear. Facebook, for example, has been criticised for allowing advertisers to target specific age groups directly.[191] Lambrecht and Tucker, who conducted one of the few empirical studies into algorithmic bias in this area, also found that algorithmically-targeted job advertisements on Facebook appeared to promote jobs in the

[189] Examples include: https://recruitmentsmart.com, https://goarya.com, https://clearfit.com, https://www.engagetalent.com, https://www.filtered.ai, https://harver.com, https://ideal.com, https://textio.com, and https://wadeandwendy.ai [all accessed on: 14/06/19].
[190] Authority of the House of Commons (2017). Algorithms in decision-making: https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf [accessed on: 14/06/19].
[191] Angwin, J., Scheiber, N., & Tobin, A. (2017). Facebook Job Ads Raise Concerns About Age Discrimination: https://www.nytimes.com/2017/12/20/business/facebook-job-ads.html [accessed on: 14/06/19].

Science, Technology, Engineering and Maths (STEM) fields to more men than women.[192] Ironically, they concluded this was mostly likely due to women being a particularly sought-after demographic within sectors which are trying to improve their diversity. This, in turn, appears to have made it more expensive to target adverts at women, and makes algorithms which are designed to prioritise cost effectiveness less likely to target them. This is an example of how algorithmic biases can develop at any stage of the process and in ways unforeseen by the developers—in this particular case, the advert was explicitly intended to be gender neutral, but was ultimately delivered in a discriminatory way.

The circumstances in which algorithms can create biased outcomes based on gender include Google advertisements generated for high-paid job ($200,000+ salary) listings. In a study that created 17,350 fake user profiles exposed to over 600,000 job listings,[193] such listings were found to be shown almost six times more often to male than to female users. The fake profiles created by the authors contained the same basic information, differing only in gender. Another recent study into the ranking of CVs for recruiters proactively searching for candidates on Indeed, Monster, and CareerBuilder found that, while there was no direct discrimination (which was precluded by the fact that gender is not collected by these websites), when researchers inferred gender (for example, from candidate names), and all other factors were controlled for, men generally had a slight advantage over women in where their CVs were ranked.[194] This, the researchers concluded, may have been due to these algorithms picking up on proxy variables (such as which universities candidates attended, and unemployment) which were weakly correlated with gender.

One of the proposed advantages in using algorithms for recruitment is that they could help remove unconscious human biases which have long existed in the recruitment process. However, most of these algorithms are suggesting courses of action based on correlations discovered in the data, rather than causal relationships.[195] Correlations tend to be considerably more unstable than causal relationships, as they may be context specific or short-lived, whereas cause and effect relationships tend to endure over time and are based on fundamental factors. Furthermore, it is also possible that advertisers are using keywords, intentionally or unwittingly, which are strongly correlated with certain age groups and protected characteristics, when targeting postings toward a certain audience. Without specific safeguards in place, these proxy keywords would go undetected. In any event, Facebook has in the past claimed that the

---

[192] Lambrecht, A., & Tucker, C. E. (2018). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018)*.
[193] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. Proceedings on privacy enhancing technologies, 2015(1), 92-112.
[194] Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018, April). Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 651). ACM.
[195] King, A. G., & Mrkonich, M. J. (2015). Big data and the risk of employment discrimination. *Okla. L. Rev.*, *68*, 555.

responsibility for managing these biases is the advertisers', suggesting that the responsibility for who puts those safeguards in place is unclear.

7.4 Case study: Recruitment sifting

In online recruitment, algorithms are now often used to filter job applications automatically based on set criteria and create a shortlist for human recruiters to then sift through manually.[196]

Unilever is a prominent example of a company which is using algorithms to screen its applicants. Every year the company processes more than 1.8 million job applications, and recruits more than 30,000.[197] They contracted Pymetrics, a specialist in AI recruitment, to create an online platform, which conducts initial screening assessments based on automated analysis of the initial written application, followed by automated analysis of 30 minute videos of applicants answering questions. The algorithm uses natural language processing and body language analysis, cross-referencing this against previous successful applicants. The company's chief of HR has said that this has saved them around 70,000 person-hours of interviewing and assessing candidates.

One of the potential issues with this kind of approach is that the data they are using to train their algorithms will sometimes reflect and perpetuate long ingrained stereotypes and assumptions about gender and race which continue to exist to this day. For example, one study found that NLP tools can learn to associate African-American names with negative sentiments, and female names with domestic work rather than professional or technical occupations.[198] Another study found similarly that systems trained on commonly used datasets learned to associate women with family and the arts and humanities, whereas men were associated with careers, maths and sciences.[199]

While these researchers note that these associations are accurate to the extent that they reflect real world trends and biases, it presents problems when companies are seeking to break from the historic patterns of employment by diversifying their workforces. For example, reports in 2018 claimed that Amazon had discontinued development of a 'sexist' machine

[196] Bogen, M., & Rieke, A. (2018). Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. *Upturn*

[197] Marr, B. (2018). The Amazing Ways How Unilever Uses Artificial Intelligence To Recruit & Train Thousands Of Employees. *Forbes*: https://www.forbes.com/sites/bernardmarr/2018/12/14/the-amazing-ways-how-unilever-uses-artificial-intelligence-to-recruit-train-thousands-of-employees/#1f84026e6274 [accessed on 17/06/19]

[198] Sutton, A., Lansdall-Welfare, T., & Cristianini, N. (2018, October). Biased Embeddings from Wild Data: Measuring, Understanding and Removing. In *International Symposium on Intelligent Data Analysis* (pp. 328-339). Springer, Cham.

[199] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

learning based tool developed at its Edinburgh office to assist internal recruitment due to concerns over gender biases it might embed.[200] The development of this tool involved training up to 500 different models to recognise up to 50,000 relevant terms on applicants' resumes, but the company found that it picked up terms male applicants used on their resumes more often when coming up with recommendations. It is unclear whether Amazon abandoned this project mainly due to these concerns, as it seems that it was generally not producing useful results.

While using AI has the potential to improve accuracy and be a cost-effective method of filtering potential employees during the recruitment process, there is a need for research into how algorithms used in this way could lead to different outcomes for particular social and ethnic groups.

7.5 Challenges and gaps

- The use of additional individual-level data, which is not strictly relevant to a particular job role, has become a growing trend in recruitment algorithms. This data, however, may also contain proxy variables which could indirectly lead to discrimination. There is a lack of governance regarding what is considered acceptable and reasonable within algorithmic processing, for example concerning what information should and needs to be used.

- There is a lack of transparency and accountability regarding the use of algorithmic decision-making in recruitment. It is difficult to know which organisations are using algorithms in their recruitment processes, what stages they are using them for, and what bias mitigation strategies they may have used. The algorithms themselves are usually difficult to study, due to both their proprietary nature, and their technical complexity. This makes it more difficult to determine possible sources of bias, as well as at which points bias prevention and mitigation efforts need to be made and which parties are responsible for making them.

- There is only limited understanding of how human recruiters interacting with algorithmic recruitment advice systems could result in new biases being introduced (which can be thought of as 'digitally mediated' bias)—for example, if human recruiters come to believe that the recommendations of automated systems are more objective and authoritative than their designers intended them to be.

[200] Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [accessed on: 14/06/19].

8. Conclusion

The literature survey conducted to prepare this report reveals that the adoption of advanced digital technologies that might be susceptible to algorithmic bias is increasing across key sectors, such as financial services, local government, crime and justice, and recruitment.

It also shows that research into algorithmic bias addresses a wide range of technical and societal factors, and that this research is closely intertwined with public debate, policy research, standardisation, and law-making efforts. This has led to a highly heterogeneous landscape of contributions, ranging from activism and public media coverage to industrial standards definition.

The present landscape demonstrates increased multi-disciplinary and cross-sector interest in the topic. But it also creates challenges in terms of coming up with a coherent roadmap for future research. There are promising signs that the transfer of knowledge from scientists, and public advocates into industry and public sector innovation is starting to happen. This link needs to be strengthened significantly to support the development of responsible practices without stifling innovation and digital transformation.

Algorithmic bias creates different challenges for each of these sectors, but there are a few cross-cutting issues that apply to all of them that give rise to a number of key recommendations:

- **Opaque systems:** there is a lack of detailed information on the internal workings of systems already in use and those that may be developed in the near future. This opacity may be caused by a lack of corporate transparency (proprietary opacity) or through the use of 'black-box' algorithmic models. This suggests a pressing need to put appropriate regulatory mechanisms in place that will allow expert scrutiny of these systems. In addition, there is a general reluctance from the companies developing these tools and the organisations using them to publish reliable details and statistics regarding their use. There is debate here about where the burden of proof and responsibility lies for making these systems more transparent. This is particularly true of areas where there is a danger that these systems might introduce or exacerbate illegal discriminatory practices.
- **More collaboration is necessary:** the scientific understanding of algorithmic bias issues is advancing, but it invariably cuts across different disciplines and needs to address a diverse range of contexts of use. As a result, we are far from convergence to a mature, unified field. Further and deeper collaboration across the computational, mathematical, legal, social, organisational, and management sciences is needed. This needs to be underpinned by research practices and policies that enable access to relevant data and

business practices, as well as wide-ranging public engagement and public policy research.

- **Regulatory and policy responses are slow:** policy making, standardisation, and documentation practices can be slow to respond to technological advancement and adoption of new business practices. Efforts to address this issue are increasing, but as scholars, politicians, and business leaders are faced with a wide range of issues around AI and ML that need to be addressed in public debate. Furthermore, it can be difficult to separate algorithmic bias considerations from many other related issues, which could further slow progress. In the interim, codes of conduct which companies can sign up to on a voluntary basis, and which individuals can hold companies accountable to even in a limited sense, may help bridge the gap, while formal legislation may ultimately prove necessary.

- **Need to look at wider "digitally mediated bias":** A critical examination of algorithmic bias issues is heavily skewed toward currently prominent AI and ML techniques, and often focuses on the opacity of the models derived by these methods. With sufficient support for research into explainability and interpretability, the scientific community might be able to develop solutions to this problem. However, the technology landscape will continue to evolve and such technical solutions will only ever provide interim solutions. A much more comprehensive understanding of the deeper problem of "digitally mediated bias" will need to be developed to come up with robust principles for the responsible development and use of future technologies.

Prediction of human behaviour is a unifying opportunity and risk

While AI and ML techniques are being adopted in a wide range of real-world applications across the world, we can discern *prediction of human behaviour* as a unifying theme of most use cases where algorithmic bias is a concern. As we observe major phenomena around the world such as the proliferation of fake news through social networks such as Facebook[201] or the Chinese "social credit" system[202], a trend toward increasingly pervasive behaviour-focused data collection and data-driven prediction of human behaviour becomes apparent.

For example, some already predict "the end of money" in a future where goods and services can be offered (or even directly allocated) to citizens based on data regarding the economic value of their work and their individual desires.[203] Machine-based optimisation of resource allocation based on global market and production conditions would ensure the right goods and services

[201] Levin, S. (2018). Facebook has a fake news 'war room'—but is it really working?: https://www.theguardian.com/technology/2018/oct/18/facebook-war-room-social-media-fake-news-politics [accessed on: 14/06/19].
[202] Kobe, N. (2019). The complicated truth about China's social credit system: https://www.wired.co.uk/article/china-social-credit-system-explained [accessed on: 14/06/19].
[203] See for example Greco, T. (2009). *The end of money and the future of civilization*. Chelsea Green Publishing; Wolman, D. (2013). *The end of money: Counterfeiters, preachers, techies, dreamers--and the coming cashless society*. Da Capo Press; Mason, P. (2016). *Postcapitalism: A guide to our future*. Macmillan.

flow to the right consumer. Similar future visions can be easily imagined in other sectors, for example in the delivery of public services, crime prevention, and labour markets, with the promise of fully automated, evidence-based, rational decision-making.

Given the increasing adoption and impact of algorithms in everyday life, it is clear that the risks to fair and equitable treatment that emanate from algorithmic bias will, if left unchecked, increase steadily. Public understanding of these issues, especially in the context of AI and machine learning, will also need to improve if citizens are to exercise their rights and hold companies and organisations to account, which may in turn require developments in education and public information campaigns. There is, however, a limit to the time and attention that we can reasonably expect even a concerned member of the public to dedicate to this subject. Ultimately, developing workable solutions to algorithmic bias requires an improved understanding amongst academics, technical experts and policymakers about the nature of the problems and the array of mitigation strategies available. We hope this literature review is a material contribution to this process.

Glossary

Algorithm: A set of precise instructions that describe how to process information, typically in order to perform a calculation or solve a problem. Algorithms have to be described in programming language to be executed on computers.

Algorithmic bias: The systematic, repeatable behaviour of an algorithm that leads to the unfair treatment of a certain group.

Artificial Intelligence (AI): An area of computer science that aims to replicate human intelligence abilities in computers. Definitions focus either on achieving human performance in complex tasks, or on mimicking the ways in which these tasks are performed by humans. In a commercial context, AI currently refers mainly to systems that use machine learning for pattern detection, prediction, human-machine dialog, and robotic control.

Attribute: A variable used as part of the description of a data sample or classifier, for example a specific pixel in a camera image, or the gender column in a spreadsheet describing employees.

Correlation: The degree to which two different statistical variables are related, often measured on the basis of how often certain values of variable A occur when certain values of variable B are observed.

Deep neural network: A neural network with many layers of nodes, each of which is capable of detecting patterns at different levels of abstraction from the previous one. Deep neural networks have been used to achieve or surpass human performance at very complex tasks. They typically require very large amounts of training data. The models learnt by deep neural nets are very hard to inspect, interpret, and explain; they currently remain largely opaque.

Direct discrimination: The process of consciously and explicitly using group membership when making decisions about an individual. Legal definitions focus on treatment of individuals with protected characteristics.

Discrimination: The process of making distinctions in the treatment of different individuals based on their actual or perceived membership to a group or social category.

Fairness: Impartial and just treatment without favouritism or discrimination in the most general sense. A complex concept that is associated, among other things, with notions of: equitable, non-discriminatory treatment in legal and administrative processes; fair distribution of wealth and other societal benefits based on concepts like social justice, solidarity, and

compassion; and appropriateness of treatment in interpersonal interaction, linked to respect and universal rights.

Machine Learning (ML): The science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions. Instead of requiring explicit programming of this model, ML algorithms identify patterns in data to develop a model that can be used to reproduce or predict the behaviour of the system they are trying to learn about. When provided with sufficient data, a machine learning algorithm can learn to make predictions or solve problems, such as identifying objects in pictures or winning at particular games.

Model (machine learning): A mathematical representation of a real-world process.[204] This may be a 'hypothesis' regarding a phenomenon described by data, that ideally provides a concise explanation of complex observations by identifying generalisable patterns and ignoring irrelevant variations.

Neural network: A network of units that compute simple numerical functions and feed their outputs into each other via weighted links. Sophisticated ML algorithms are capable of adapting these weights in ways that allow a large enough network to capture any complex mathematical function using sufficiently large amounts of training data.

Protected characteristics: Attributes of individuals explicitly protected by anti-discrimination law. In the UK these are legally defined under the Equality Act 2010, and cover age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, and sex.

[204] Bhattacharjee, J. (2017). Some key machine learning definitions. *NineLeaps*, available at: https://medium.com/technology-nineleaps/some-key-machine-learning-definitions-b524eb6cb48 [accessed on: 11/07/19].